Non-linear bias of cosmological halo formation in the early universe

Article (Published Version)

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **450,** 1486–1502 (2015) doi:10.1093/mnras/stv704

Non-linear bias of cosmological halo formation in the early universe

Kyungjin Ahn,[1⋆] Ilian T. Iliev,[2] Paul R. Shapiro[3] and Chaichalit Srisawat[2]

[1]*Department of Earth Sciences, Chosun University, Gwangju 501-759, Korea*
[2]*Astronomy Centre, Department of Physics and Astronomy, Pevensey II Building, University of Sussex, Falmer, Brighton BN1 9QH, UK*
[3]*Department of Astronomy, University of Texas, Austin, TX 78712-1083, USA*

Accepted 2015 March 27. Received 2015 March 25; in original form 2014 July 9

ABSTRACT

We present estimates of the non-linear bias of cosmological halo formation, spanning a wide range in the halo mass from $\sim 10^5$ to $\sim 10^{12}\,M_\odot$, based upon both a suite of high-resolution cosmological *N*-body simulations and theoretical predictions. The halo bias is expressed in terms of the mean bias and stochasticity as a function of local overdensity (δ), under different filtering scales, which is realized as the density of individual cells in uniform grids. The sampled overdensities span a range wide enough to provide the fully non-linear bias effect on the formation of haloes. A strong correlation between δ and halo population overdensity δ_h is found, along with sizable stochasticity. We find that the empirical mean halo bias matches, with good accuracy, the prediction by the peak-background split method based on the excursion set formalism, as long as the empirical, globally averaged halo mass function is used. Consequently, this bias formalism is insensitive to uncertainties caused by varying halo-identification schemes, and can be applied generically. We also find that the probability distribution function of biased halo numbers has wider distribution than the pure Poisson shot noise, which is attributed to the sub-cell-scale halo correlation. We explicitly calculate this correlation function and show that both overdense and underdense regions have positive correlation, leading to stochasticity larger than the Poisson shot noise in the range of haloes and halo-collapse epochs we study.

Key words: galaxies: haloes – cosmology: theory – large-scale structure of Universe.

1 INTRODUCTION

In the standard scenario of cosmological structure formation, cosmological haloes are the features of the cosmic web of highest overdensity in approximate virial equilibrium, that result from the non-linear amplification of initially linear, Gaussian-random density fluctuations by gravitational instability. Galaxies and haloes, however, are not unbiased tracers of the underlying density distribution, and thus understanding this 'bias' effect is crucial to extract cosmological information from the data of galaxy surveys, for example.

The idea that galaxy bias (from this point on, we will sometimes denote dark matter haloes loosely by 'galaxies' to reflect the original ideas of associating galaxies purely by high-density peaks without resorting to hydrodynamical cooling mechanism) exists and can be calculated from the statistics of Gaussian-random initial density fields was pioneered by Kaiser (1984). Bardeen et al. (1986) extended this idea to take a full account of the Gaussian-random density field in a cosmological context, to understand how haloes grow out of this random field and cluster spatially. In the meantime, Press & Schechter (1974, PS hereafter) associated cosmological haloes (or galaxies) as high-density peaks and estimated halo mass function, and this PS formalism was recounted more rigorously by Bond et al. (1991) through their excursion set formalism (sometimes called the extended PS formalism), where they showed that cloud-in-cloud effect explains the fudge multiplicity factor 2 in the PS mass function. All these ideas form the backbone of the peak-background split scheme for calculating the galaxy bias by Cole & Kaiser (1989), which bears the idea that haloes (peaks) are more typically formed in high-density regions. Mo & White (1996, MW hereafter) calculated fully non-linear bias combining the peak-background split scheme with the spherical top-hat collapse model under the extended Press–Schechter formalism, and also calculated the useful 'linear bias parameter' in the linear regime. The peak-background split scheme may not give a perfectly accurate prediction of *N*-body simulation results (e.g. MW; Manera, Sheth & Scoccimarro 2010), which is usually attributed to the discrepancy between the PS mass function and the *N*-body halo mass function at low- and high-mass ends (e.g. Sheth & Tormen 1999, ST hereafter; Jenkins et al. 2001). This discrepancy stimulated better-fitting functional forms (e.g. ST; Jenkins et al. 2001; Warren et al. 2006;

⋆ E-mail: kjahn@chosun.ac.kr

Lukić et al. 2007; Reed et al. 2007; Lim & Lee 2013; Watson et al. 2014). Barkana & Loeb (2004, BL hereafter) then developed a hybrid scheme of combining ST mass function and the linear bias parameter derived from the extended Press–Schechter formalism and showed that this fitted the linear *N*-body halo bias better than MW prediction.

Bias can of course have stochasticity, which was formulated theoretically by Dekel & Lahav (1999, DL henceforth): haloes sampled inside a suite of Eulerian cells of a given density, or count-in-cell haloes, are expected to deviate from purely Poisson distribution, if there is either correlation or anticorrelation of haloes at sub-cell scales which then result in variance of the number of haloes ($\sigma^2(N)$) larger or smaller than Poissonian value, respectively (e.g. Peebles 1993; see also Section 3.4). Somerville et al. (2001) compared the prediction by DL to *N*-body simulation results, and based on the observed $\sigma^2(N)$ they concluded that haloes are usually correlated in overdense regions and anticorrelated in underdense regions (we will however contradict this claim in Section 4.3). Later work found that haloes usually show variance larger than the Poissonian value (e.g. Neyrinck et al. 2014 find that haloes of mass $10^{10\text{–}11}\,\mathrm{M}_\odot$ show this 'super-Poissonian' distribution under $2\,h^{-1}$ Mpc cells), which are well fitted by the functional distributions suggested by Saslaw & Hamilton (1984) and Sheth (1995).

A useful application of the non-linear halo bias prescription is to create mock halo catalogues in a large scale for either cosmology or astrophysics. While mock galaxy catalogues can be created by schemes based on quasi-linear perturbation theory, such as PINOCCHIO (Monaco et al. 2002, 2013) and PTHALOES (Scoccimarro & Sheth 2002; Manera et al. 2013), they are usually limited to the scales under which density perturbation remains quasi-linear at most. This limitation can be overcome by non-linear halo bias schemes, as in Kitaura, Yepes & Prada (2014) who prove the concept by generating halo catalogues which are statistically consistent with *N*-body halo catalogues, suited for probing the baryon acoustic oscillation feature by surveys such as the Baryon Oscillation Spectroscopic Survey. We intend to achieve a similar goal in the long run, but with a bias scheme that is fully non-linear and is applicable regardless of the halo mass, the filtering scale and the redshift. Because we will calculate the bias parameter theoretically, our scheme will mitigate the need to find an empirical fitting formula as done in e.g. Kitaura et al. (2014).

A similar formalism can also be applied to astrophysical problems. Understanding the halo bias is crucial e.g. in the study of cosmic reionization, due to the very large dynamic gap between the very small galaxies believed to be the main drivers of reionization (see e.g. Ciardi & Ferrara 2005, for a review) and the large characteristic scales of the reionization patchiness (Friedrich et al. 2011; Iliev et al. 2014). BL used a hybrid halo bias scheme to study the fluctuation of the 21 cm background from the fluctuating halo distribution during the epoch of reionization (EoR). Fast seminumerical simulators of reionization (Zahn et al. 2007; Santos et al. 2008; Alvarez et al. 2009; Mesinger, Furlanetto & Cen 2011), whose basis was formulated by Furlanetto, Zaldarriaga & Hernquist (2004) and Furlanetto & Oh (2005) to replace the time-consuming ray-tracing by a faster excursion set formalism, make use of a similar formalism to seed haloes in a coarse-grained density field.

We have indeed applied this formalism to a simulation of cosmic reionization, by which we could span the full dynamic range of haloes hosting radiation sources. Cosmic reionization is believed to occur very inhomogeneously with large H II regions, whose sizes show a wide distribution peaked at ∼20 comoving Mpc before completion if roughly put. Therefore it is necessary to use a large box in order to simulate the reionization process in a statistically reliable way. This requirement, however, limits the ability of the simulation to resolve 'minihaloes' which are believed to host Population III stars, and allows the simulation to only resolve the more massive kind, or 'atomic-cooling haloes'. Indeed most reionization simulations in large boxes used to implement atomic-cooling haloes only, while this may underestimate the photon budget in the early stage of reionization. In a large-scale (box size of $114\,h^{-1}$ Mpc comoving) simulation of cosmic reionization (with ray-tracing method), Ahn et al. (2012) used the conditional halo bias found in Section 4.2.1 of this paper to include minihaloes, which could not otherwise have been realized due to numerical resolution. This way, they could span the full dynamic range of haloes – both minihaloes and atomic-cooling haloes – responsible for emitting hydrogen-ionizing and H_2-dissociation radiation, and observed that the reionization process is extended further in time to comply better with several observational constraints.

On much larger scales (box size of $425\,h^{-1}$ Mpc comoving) the same technique was used to perform the largest volume, ray-tracing simulations of cosmic reionization to date, presented in Iliev et al. (2014) and further explored in Datta et al. (2012), Park et al. (2013) and Shapiro et al. (2013). This used the results in Section 4.2.2 to include the unresolved low-mass atomic cooling haloes ($M = 10^8$–$10^9\,\mathrm{M}_\odot$). Another prospective application is in exploring the effects of primordial non-Gaussianity on halo bias, which is an active area of research (e.g. Dalal et al. 2008; Adshead et al. 2012; D'Aloisio et al. 2013), and which also leads to ionization bias (e.g. Joudaki et al. 2011; D'Aloisio et al. 2013) detectable by 21 cm observations (e.g. Mao et al. 2013).

In this paper, we examine and compare the non-linear halo bias from both our suite of cosmological *N*-body simulations suited for the study of haloes responsible for EoR and a semi-analytical, fully non-linear peak-background split scheme. This theoretical scheme is a hybrid scheme similar to the one by BL, but also differs as we combine the empirical (mean) halo mass function to the bias factor and extend it to the fully non-linear regime in a non-perturbative way. Through this, we investigate whether the bias factor can be purely based upon the excursion set formalism and separated cleanly from the mass function, which bears uncertainty due to its strong dependence on specific halo-identification schemes. We also study the stochasticity of halo bias from these simulations and examine whether they are purely Poissonian or not, which has been investigated recently to conclude that haloes in some mass range indeed have super-Poissonian distribution (Baldauf et al. 2013; Neyrinck et al. 2014). Towards this, we calculate the two-point halo correlation function and quantify its contribution to stochasticity in addition to the Poisson noise. While our paper is focused on the range of haloes responsible for cosmic reionization at $z \gtrsim 6$, and therefore it can be used readily in the study of EoR, our formalism should be applicable in more generic cases.

This paper is organized as follows. In Section 2, we briefly describe our *N*-body simulation. In Section 3, we describe the theoretical scheme for the non-linear halo bias, which combines the peak-background split scheme (Sections 3.1 and 3.2) with the empirical *N*-body halo mass function (Section 3.3), and also describe the stochasticity and various quantities related (Section 3.4). We then describe our results in Section 4, first on the mean halo mass function (Section 4.1), then on the mean bias (Section 4.2) and on the stochasticity (Section 4.3). We further investigate the validity of the usual linear bias approximation in Section 4.4. We conclude our paper in Section 5, together with a schematic layout of our bias prescription towards generating mock halo catalogues.

Table 1. *N*-body simulation parameters. Background cosmology is based on the *WMAP* 5 yr results: $\Omega_{\rm m} = 0.27$, $\Omega_\Lambda = 0.73$, $\Omega_{\rm b} = 0.044$, $h = 0.7$, $\sigma_8 = 0.8$ and $n_{\rm s} = 0.96$.

Simulation	Box size	$N_{\rm particle}$	Mesh	Spatial resolution	$m_{\rm particle}$	$M_{\rm halo,\,min}$
S1	$6.3\,h^{-1}$ Mpc	1728^3	3456^3	$182\,h^{-1}$ pc	$5.19 \times 10^3\,{\rm M}_\odot$	$1.04 \times 10^5\,{\rm M}_\odot$
M1	$20\,h^{-1}$ Mpc	5488^3	$10\,976^3$	$182\,h^{-1}$ pc	$5.19 \times 10^3\,{\rm M}_\odot$	$1.04 \times 10^5\,{\rm M}_\odot$
B1	$114\,h^{-1}$ Mpc	3072^3	6144^3	$1.86\,h^{-1}$ kpc	$5.47 \times 10^6\,{\rm M}_\odot$	$1.09 \times 10^8\,{\rm M}_\odot$

2 SIMULATIONS

The data used in this work are based on a suite of large simulations, most of which were previously presented in Watson et al. (2014). They were performed using the CUBEP^{3}M code, a high-performance, publicly available, cosmological *N*-body code based on particle–particle–particle–mesh (P^{3}M) scheme (for detailed code description and tests see Harnois-Déraps et al. 2013). For memory efficiency and speed the code uses two-level grid for computing the long-range gravity forces using a particle–mesh method and adds the local direct particle–particle forces at small scales. CUBEP^{3}M is a massively parallel, hybrid (using MPI and OpenMP) code, scaling well up to tens of thousands of computing cores. It has been extensively tested and run on a wide variety of parallel platforms.

Our complete simulation suite, listed in Table 1, includes volumes between $6.3\,h^{-1}$ Mpc and $114\,h^{-1}$ Mpc per side and between 1728^3 and 5488^3 particles, thereby covering a large dynamic range, with particle masses ranging from 5.2×10^3 to $5.5 \times 10^6\,{\rm M}_\odot$ and force smoothing lengths between 182 pc and $1.86\,h^{-1}$ kpc. The smaller-volume, high-resolution simulations with boxes up to $20\,h^{-1}$ Mpc per side resolve (with 20 particles or more) dark matter haloes with mass $10^5\,{\rm M}_\odot$ and above, the expected hosts of the first stars. In contrast, the larger volume, $114\,h^{-1}$ Mpc only resolves haloes with mass $10^8\,{\rm M}_\odot$ (with 20 particles) and larger, but samples the statistics of rare haloes much better due to its larger volume.

We locate the collapsed haloes at runtime, using the CPMSO spherical overdensity method (Harnois-Déraps et al. 2013; Watson et al. 2014) with overdensity with respect to the mean of 178, suitable for the high redshifts considered here. This is done by first interpolating the particles on to a fine grid (with number of cells per dimension twice the number of particles) using the cloud-in-cell approximation. Local density peaks (with density at least 100 times the average) are located and spherical shells are expanded around each peak until the threshold overdensity is crossed. The resulting object is then marked as a halo (objects with less than 20 particles are discarded as they cannot be reliably identified). The halo centre position is calculated more precisely by quadratic interpolation within the cell and the particles within the halo virial radius are identified and then the halo properties, e.g. mass, velocity dispersion, centre of mass, angular momentum, radius, etc., are calculated and saved in the halo catalogue.

3 THEORY

Formation of cosmological haloes is strongly correlated with their larger scale density environment. The excursion set formalism (Bond et al. 1991) gives a quantitative description of this biased halo formation in terms of the conditional halo mass function $\mathrm{d}n/\mathrm{d}M\,(M;\delta)$, where $\delta \equiv (\rho - \bar{\rho})/\bar{\rho}$ is the overdensity of the local environment. This description is called the peak-background split, where haloes are considered as the high-density 'peaks' that are placed on large-scale density 'background'. In the linear regime where $\delta \ll 1$, this yields the linear bias parameter which has been used extensively in cosmology (MW).

In this section we introduce a formalism which is intended to describe the local non-linear bias in a non-perturbative way, based mostly on the formalism by MW and the idea of BL. Therefore, we revisit previous theoretical work, and at the same time describe modifications we made in this section. We will then compare the prediction from this formalism to the *N*-body data results in Section 4. We will occasionally add subscript 'L' to Lagrangian quantities, when otherwise these may be confused with Eulerian ones.

3.1 Biased halo mass function in Lagrangian volume

It is shown in the excursion set formalism that distribution of linear overdensity δ in the initially Gaussian-random matter density field $\rho_{\rm L}$ filtered with a 'sharp *k*-space filter',

$$\rho_{\rm L}(\boldsymbol{r},\,R_{\rm f,L}) = \int \mathrm{d}^3 r'\, W_{\rm K}(\boldsymbol{r}-\boldsymbol{r}';\,R_{\rm f,L})\rho_{\rm L}(\boldsymbol{r}',0),$$

$$\rho_{\rm L}(k,\,R_{\rm f,L}) = \tilde{W}_{\rm K}(k;\,R_{\rm f,L})\rho_{\rm L}(\boldsymbol{k},0), \qquad (1)$$

where the window function $W_{\rm K}(r;R_{\rm f,\,L})$ is the Fourier transform of sharp *k*-space filter $\tilde{W}_{\rm K}(k;\,R_{\rm f,L}) \equiv \Theta(1 - k\,R_{\rm f,L})$, still follows Gaussian distribution (Bond et al. 1991). This is obviously true even in the density field linearly extrapolated to the observing redshift with the linear growing factor. This way, one can use the linearly extrapolated density field and the appropriate halo-collapse criterion to predict halo population at any filter scale and redshift.

In this formalism the unconditional, globally averaged differential halo number density (mass function) is given by the Press–Schechter formula (PS)

$$\left(\frac{\mathrm{d}n}{\mathrm{d}M}\right)_{\rm PS}(M) = \left(\frac{\mathrm{d}n}{\mathrm{d}M}\right)_{\rm PS}(\sigma^2_{M,\rm L};\delta_{\rm c})$$
$$= -\frac{1}{\sqrt{2\pi}}\frac{\mathrm{d}\sigma^2_{M,\rm L}}{\mathrm{d}M}\frac{\bar{\rho}_0}{M}\frac{\nu}{\sigma^2_{M,\rm L}}\exp\left[-\frac{\nu^2}{2}\right], \qquad (2)$$

where $\sigma^2_{M,\rm L}$ is the variance of Gaussian distribution of the density field (linearly extrapolated to the present) filtered in real space in spheres with radius $R_{\rm f,\,L}$, $\bar{\rho}_0$ is the present matter density, $\nu \equiv \delta_{\rm c}/(D(z)\sigma_{M,\,\rm L})$ (with the linear growth factor $D(z)$ in ΛCDM universe) is the ratio of critical overdensity $\delta_{\rm c} = 1.686$[1] to $\sigma_{M,\,\rm L}(z) = D(z)\sigma_{M,\,\rm L}$ and $R_{\rm f,\,L}$ is the length-scale usually associated [2] with the halo mass M by

$$M = M(R_{\rm f,L}) = \bar{\rho}_0\frac{4\pi}{3}R^3_{\rm f,L}. \qquad (3)$$

[1] We neglect the very weak redshift dependence of $\delta_{\rm c}$ in ΛCDM in our study, while for $z \lesssim 4$ one should implement its redshift dependence.
[2] Rigorously speaking, we cannot associate such a well-defined mass M with $R_{\rm f,\,L}$ in the case of sharp k filtering (e.g. Bond et al. 1991). However, we adopt this definition for simplicity.

Equation (2) is unconditional in a sense that this represents the average halo distribution in the universe.

The excursion set formalism also predicts halo population inside a region with given mean overdensity and size. Sharp k-space filtering allows one to write this in terms of conditional probability analytically, because wavemodes at different filter scales are linearly independent. The barrier crossing and variance under given density environment, which will be denoted by a 'cell', is measured from the new origin $\delta_{\rm lin}$ (throughout this paper, unless specified differently, we denote the full, non-linear overdensity of a cell by δ for simplicity) and $\sigma_{\rm cell,\,L}$, which are linearly extrapolated density of the cell and variance corresponding to the Lagrangian cell size $R_{\rm cell,\,L}$, respectively. When a Eulerian cell has a comoving volume $V_{\rm cell}$ and non-linear overdensity δ at some redshift, $R_{\rm cell,\,L}$ can be obtained from

$$M_{\rm cell} = \bar{\rho}_0 V_{\rm cell}\,(1+\delta) = \bar{\rho}_0 \frac{4\pi}{3} R^3_{\rm cell,L}. \tag{4}$$

According to the well-known excursion-set formalism (Bond et al. 1991), the differential halo number density (halo mass function) inside a Lagrangian region with $\delta_{\rm lin}$ (linearly extrapolated to redshift z) and $R_{\rm cell,\,L}$ is then given by a conditional mass function

$$\begin{aligned}\left(\frac{{\rm d}n}{{\rm d}M}\right)^{\rm L}_{\rm PS,b}(M|\delta_{\rm lin}) &\equiv \left(\frac{{\rm d}n}{{\rm d}M}\right)_{\rm PS}(\sigma^2_{M,\rm L};\,\delta_{\rm c}|\sigma^2_{\rm cell,L};\,\delta_{\rm lin}) \\ &= \left(\frac{{\rm d}n}{{\rm d}M}\right)_{\rm PS}(\sigma^2_{M,\rm L}-\sigma^2_{\rm cell,L};\,\delta_{\rm c}-\delta_{\rm lin}) \\ &= -\frac{1}{\sqrt{2\pi}}\frac{{\rm d}\sigma^2_{M,\rm L}}{{\rm d}M}\frac{\bar{\rho}_0}{M}\frac{(\delta_{\rm c}-\delta_{\rm lin})/D(z)}{\left(\sigma^2_{M,\rm L}-\sigma^2_{\rm cell,L}\right)^{3/2}} \\ &\quad\times\exp\left[-\frac{(\delta_{\rm c}-\delta_{\rm lin})^2}{2D^2(z)\left(\sigma^2_{M,\rm L}-\sigma^2_{\rm cell,L}\right)}\right],\end{aligned} \tag{5}$$

which takes the same form as equation (2) but with $\sigma^2_{M,\rm L}$ and $\delta_{\rm c}$ replaced by $\sigma^2_{M,\rm L}-\sigma^2_{\rm cell,L}$ and $\delta_{\rm c}-\delta_{\rm lin}$, respectively. Here, $\sigma_{\rm cell,L}\equiv\sigma_{M_{\rm cell},\rm L}$. This defines the Lagrangian overabundance of haloes of mass M,

$$\delta^{\rm L}_{\rm h}(M|\delta_{\rm lin}) \equiv \left(\frac{{\rm d}n}{{\rm d}M}\right)^{\rm L}_{\rm PS,b}(M|\delta_{\rm lin}) \Big/ \left(\frac{{\rm d}n}{{\rm d}M}\right)_{\rm PS}(M) - 1. \tag{6}$$

Note that two important factors should be considered in order to generalize equation (5). First, in the non-linear regime where $\delta_{\rm lin}\sim 1$, one should match the non-linear δ to the linear $\delta_{\rm lin}$ to use equation (5), because this is based on the linear theory. Secondly, $({\rm d}n/{\rm d}M)^{\rm L}_{\rm PS,b}$ and $\delta^{\rm L}_{\rm h}$ should be converted into the corresponding Eulerian mass function and Eulerian halo overabundance, respectively, because Eulerian quantities are of much more practical use than Lagrangian quantities. This conversion will be described in Section 3.2.

3.2 Non-linear background and biased haloes mass function in Eulerian volume

The quantities $({\rm d}n/{\rm d}M)^{\rm L}_{\rm PS,b}$ and $\delta^{\rm L}_{\rm h}$ in equations (5) and (6) are derived assuming that density grows linearly with the linear growth factor and are defined in the Lagrangian volume. In reality, growth of density perturbations is non-linear in general, and this also yields large difference between the Lagrangian and Eulerian volumes.

Therefore, we first need to map non-linear overdensity δ to linear overdensity $\delta_{\rm lin}$. We use the mapping scheme based on the top-hat collapse model, which has also been used by MW, where δ, which is non-linear in general, is linked to $\delta_{\rm lin}$ in a parametric form of θ as follows:

$$\delta = \left(\frac{10\,\delta_{\rm lin}}{3(1-\cos\theta)}\right)^3 - 1, \quad \delta_{\rm lin} = \frac{3\times 6^{2/3}}{20}(\theta-\sin\theta)^{2/3}, \tag{7}$$

if $\delta > 0$. Similarly, if $\delta < 0$,

$$\delta = \left(\frac{10\,\delta_{\rm lin}}{3(\cos h\theta - 1)}\right)^3 - 1, \quad \delta_{\rm lin} = \frac{3\times 6^{2/3}}{20}(\sin h\theta-\theta)^{2/3}. \tag{8}$$

Note that δ increases monotonically as $\delta_{\rm lin}$ increases, such that there exists one-to-one mapping.

We also need to consider the change of Lagrangian volume by multiplying the ratio of Lagrangian volume to the Eulerian volume to obtain the correct Eulerian number density, which yields the final form:

$$\left(\frac{{\rm d}n}{{\rm d}M}\right)_{\rm PS,b} = \left(\frac{{\rm d}n}{{\rm d}M}\right)^{\rm L}_{\rm PS,b}(1+\delta). \tag{9}$$

By taking further approximation that $\delta_{\rm c}\gg\delta$ and $\sigma_M\gg\sigma_{\rm cell}$ MW find a useful linear relation between $\delta_{\rm h}$ and $\delta_{\rm cell}$. This approximation implies that total mass contained in haloes inside a cell is much smaller than the total mass of the cell. However, this approximation is not always valid at high resolution because some cells in our density field, depending on the choice of the cell size, may achieve very high overdensity $\delta_{\rm cell}$ such that $\delta_{\rm c}\gtrsim\delta$ and $\sigma_M\gtrsim\sigma_{\rm cell}$. Therefore, we just use equation (9) in its general form, which allows for non-linear relation between $\delta_{\rm h}$ and $\delta_{\rm cell}$.

3.3 Non-linear bias and hybrid conditional mass function

Before proceeding, let us define the mean conditional bias function $b(\delta)$ (MW; DL):

$$b(\delta) \equiv \frac{\langle\delta_{\rm h}(M|\delta)\rangle_{\delta_{\rm h}|\delta}}{\delta}, \tag{10}$$

where $\delta_{\rm h}(M|\delta)$ is the conditional, Eulerian halo overabundance, and the seemingly repetitive definition of the average is to clarify the fact that the average is taken *only* over the cells with the given δ, following the notation from equations 3 and 4 of DL, which is different from the average over all cells regardless of δ, or $\langle\ \rangle$. This average takes the following integral form for any conditional function of $\delta_{\rm h}$ under a given δ, $f(\delta_{\rm h})|\delta$:

$$[f(\delta_{\rm h}|\delta)] \equiv \langle f(\delta_{\rm h}|\delta)\rangle_{\delta_{\rm h}|\delta} \equiv \int {\rm d}\delta_{\rm h}\,P(\delta_{\rm h}|\delta)f(\delta_{\rm h}), \tag{11}$$

where $P(\delta_{\rm h}|\delta)$ is the conditional probability for a cell with δ to have $\delta_{\rm h}$ as the halo overabundance inside it (DL), and only those cells with given δ are included in the integration. To distinguish the *conditional* averaging from the normal averaging $\langle f\rangle$, we denote the former by a simple notation, $[f]$, in which the dependence on δ is assumed implicitly. Equation (11) is equivalent to equation 5 in DL.

Equations (2), (5) and (9) naturally determine by how much the local halo mass function is modified. The Eulerian overabundance of haloes is then given by

$$\begin{aligned}[\delta_{\rm h}(M|\delta)] &= \frac{\left(\frac{{\rm d}n}{{\rm d}M}\right)_{\rm PS,b}(M|\delta)}{\left(\frac{{\rm d}n}{{\rm d}M}\right)_{\rm PS}(M)} - 1 \\ &= \frac{\left(\frac{{\rm d}n}{{\rm d}M}\right)_{\rm PS}(\sigma^2_{M,\rm L};\,\delta_{\rm c}|\sigma^2_{\rm cell,L};\,\delta_{\rm lin})}{\left(\frac{{\rm d}n}{{\rm d}M}\right)_{\rm PS}(\sigma^2_{M,\rm L};\,\delta_{\rm c})}(1+\delta) - 1,\end{aligned} \tag{12}$$

which is equivalent to equation 19 of MW. The bias function b becomes independent of δ in the linear regime where $\delta_c \gg |\delta| \simeq |\delta_0|$ and $\sigma^2_{M,\rm L} \gg \sigma^2_{\rm cell,L}$, and is given as a function of ν alone, at any given z:

$$b_{\rm lin}(\delta) = 1 + \frac{\nu^2 - 1}{\delta_c(z)} \tag{13}$$

(MW). $b_{\rm lin}$ is referred to as the linear bias parameter. We will test the applicability of this approximation in Sections 4.2 and 4.4.

The relation between $\delta_{\rm h}$ and δ is generally non-linear, and therefore equation (13) is of limited use for our purposes. Even in the linear regime where $|\delta| \ll 1$, using equation (13) may be problematic because the other condition $\sigma^2_M \gg \sigma^2_{\rm cell}$ is not valid in general and then the exponential term in equation (5) cannot be approximated further. For example, for minihaloes of $M \geq 10^5\,{\rm M}_\odot$, we have $\sigma^2_M \leq 70.6$, while cells we study here have masses (when $\delta = 0$) as low as $3.5 \times 10^8\,{\rm M}_\odot$ (in both 6.3 and $20\,h^{-1}$ Mpc boxes), which corresponds to $\sigma^2_{\rm cell} = 24.0$.

One may naively expect that $({\rm d}n/{\rm d}M)_{\rm PS,\,b}$ gives the correct analytical estimate for the biased halo mass function. However, it is well known that the unconditional PS mass function, $({\rm d}n/{\rm d}M)_{\rm PS}$, is a poor fit to the empirical halo mass function derived from N-body simulations, in general, depending on the range of mass – especially so for rare haloes – and redshift (e.g. Jenkins et al. 2001). It is thus reasonable to expect that $({\rm d}n/{\rm d}M)_{\rm PS,\,b}$ will also become a poor fit to the biased N-body halo mass function.

We therefore adopt a hybrid approach, first introduced by BL, to predict the conditional mass function (or bias) by combining $\delta_{\rm h}(\delta)$ (or equivalently $b(\delta)$) as in equation (12), derived from the excursion set formalism, with the unconditional mass function ${\rm d}n/{\rm d}M$, which we choose independently. This approach is somewhat advantageous over Sheth & Tormen (2002) and PS, for example, because $\delta_{\rm h}(\delta)$ or $b(\delta)$ is almost independent of how haloes are identified (MW) and thus the unconditional mass function can be found empirically for any arbitrarily identified N-body haloes. We can then expect that when such an empirical mass function ${\rm d}n/{\rm d}M$ is combined with equation (12), the resulting mass function may be a better fit to the actual biased halo mass function $({\rm d}n/{\rm d}M)_{\rm b}$.

In contrast to BL, who choose the well-known PS and Sheth–Tormen (ST) mass functions, we choose three mass functions: PS, ST and the empirical fit to our N-body data. The reason for using the empirical (unconditional) mass function is because (1) both PS and ST mass functions are known to be poor fits to very rare haloes (see discussion in Watson et al. 2014 and references therein) and for the redshift and halo mass range of interest here all haloes are rare and (2) we want a prescription which is independent of the systematic uncertainties of the unconditional mass function due to the varying halo-identification schemes. The conditional PS bias trivially reduces to $({\rm d}n/{\rm d}M)_{\rm PS,\,b}$, while in the other two cases, the unconditional ST $({\rm d}n/{\rm d}M)_{\rm ST}$ and the empirical fit $({\rm d}n/{\rm d}M)_{N-{\rm body}}$ are both simply multiplied by $1 + \delta_{\rm h}$ to produce

$$\left(\frac{{\rm d}n}{{\rm d}M}\right)_{\rm ST,b} = \{1 + \delta_{\rm h}(\delta)\}\left(\frac{{\rm d}n}{{\rm d}M}\right)_{\rm ST} = \{1 + b(\delta)\delta\}\left(\frac{{\rm d}n}{{\rm d}M}\right)_{\rm ST} \tag{14}$$

and

$$\left(\frac{{\rm d}n}{{\rm d}M}\right)_{N-{\rm body,b}} = \{1 + \delta_{\rm h}(\delta)\}\left(\frac{{\rm d}n}{{\rm d}M}\right)_{N-{\rm body}} = \{1 + b(\delta)\delta\}\left(\frac{{\rm d}n}{{\rm d}M}\right)_{N-{\rm body}}, \tag{15}$$

where $\delta_{\rm h}(\delta)$ is given by equation (12). It is important to note that even when $\delta = 0$, $(1 + \delta_{\rm h}) \neq 1$ in general. In order to illustrate this, let us consider the limiting case of very rare haloes such that $\nu \gg 1$. Such haloes will most likely form at very high density regions – or more explicitly, high-density cells with some fixed Eulerian volume – with $\delta \gg 0$. In this case, $(1 + \delta_{\rm h}) \to 0$ or $b(\delta)\delta \to -1$ as $\delta \to 0$, and thus a simple linear relation $\delta_{\rm h} \propto \delta$, which yields $(1 + \delta_{\rm h}) \to 1$ as $\delta \to 0$, inevitably fails in estimating the bias correctly even in the linear regime. More detailed discussion of this aspect is in Section 4.4.

Finally, the fraction of halo mass to cell mass, or the collapsed fraction, is given by

$$f_{\rm c,b}(M_{\rm min}, M_{\rm max}) \equiv f_{\rm c}\left(M_{\rm min}, M_{\rm max}|\sigma^2_{\rm cell};\,\delta\right) = \frac{\int_{M_{\rm min}}^{M_{\rm max}} \left(\frac{{\rm d}n}{{\rm d}M}\right)_{\rm b} M{\rm d}M}{\rho_0(1+\delta)} = \frac{\int_{M_{\rm min}}^{M_{\rm max}} \left(\frac{{\rm d}n}{{\rm d}M}\right)^{\rm L}_{\rm b} M{\rm d}M}{\rho_0}, \tag{16}$$

which is naturally expressed in Lagrangian quantities, because both masses inhabit the same Lagrangian region. Here once again, $({\rm d}n/{\rm d}M)_{\rm b}$ can be based on either the PS mass function, the ST mass function or the empirical fit to simulations.

3.4 Expected stochasticity and renormalization

We have so far described the mean conditional mass function. In reality, the observed correlation should exhibit stochasticity as well, because structure forms out of a random density field. In addition, when haloes of our interest are rare, not all the cells with given δ will contain such haloes, giving rise to Poisson fluctuations. However, we will soon see that the stochasticity should differ from pure Poissonian distribution. Here we consider only the local stochasticity and postpone the analysis of multipoint correlation and corresponding statistics to a future paper.

Because the conditional mass function has a stochastic element, the total number of haloes inside cells with given overdensity δ and Eulerian volume $V_{\rm cell}$ would show a scatter around the mean value. For the total number of haloes in a mass bin $M = [M_{\rm min}, M_{\rm max}]$,

$$N(M_{\rm min}, M_{\rm max}|\delta, V_{\rm cell}) \equiv V_{\rm cell}\int_{M_{\rm min}}^{M_{\rm max}} {\rm d}M\left(\frac{{\rm d}n}{{\rm d}M}\right)_{\rm o,cell}, \tag{17}$$

over different cells with the same $V_{\rm cell}$ and δ and where $({\rm d}n/{\rm d}M)_{\rm o,\,cell}$ is the observed halo mass function inside each cell, one would naively expect that the probability distribution function (PDF) of N will obey the Poisson statistics:

$$P_{\rm cell}(N) \equiv P(N|\delta, V_{\rm cell}) \to \frac{{\rm e}^{-[N]}\,[N]^N}{N!}, \tag{18}$$

where the average is again taken only over the cells with given δ such that $[N] = \langle N\rangle_{\delta_{\rm h}|\delta} = \langle N(M_{\rm min}, M_{\rm max}|\delta, V_{\rm cell})\rangle_{\delta_{\rm h}|\delta}$. If so, both the conditional mean and conditional variance of N would become identical to $[N]$. However, if correlation of haloes at sub-cell length-scale exists, there occurs an additional variance – either positive or negative – in N (Peebles 1993; DL):

$$\Delta_{\rm scc}(\delta) = \left(\frac{[N]}{V_{\rm cell}}\right)^2\int^{V_{\rm cell}} {\rm d}V_1{\rm d}V_2\,\overline{\xi_{12}}(\delta), \tag{19}$$

where 'scc' denotes sub-cell correlation such that the integration is taken inside a cell and the *conditional* sub-cell two-point correlation

function $\overline{\xi_{12}}(\delta)$ is defined by

$$[N_1 N_2] = \left(\frac{[N]}{V_{\rm cell}}\right)^2 {\rm d}V_1 {\rm d}V_2 \left\{1 + \overline{\xi_{12}}(\delta)\right\}, \quad (20)$$

where 1 and 2 denote two different sub-cell positions inside the same cell and N_1 and N_2 are number of haloes in each sub-cell. $\overline{\xi_{12}}(\delta)$ should not be confused with the global sub-cell correlation function ξ_{12}, defined by

$$\langle N_1 N_2 \rangle = \left(\frac{\langle N \rangle}{V_{\rm cell}}\right)^2 {\rm d}V_1 {\rm d}V_2 \left\{1 + \xi_{12}\right\}. \quad (21)$$

Note that equations (19) and (20) are restricted only to cells with given δ, which are direct applications of equations 7.66 and 7.63 in Peebles (1993), respectively. While these equations were originally intended for unconditional quantities in Peebles (1993), applying these to conditional quantities is trivially achieved by replacing the global average $\langle\ \rangle$ with the conditional average $[\ ]$. This is easily justified by the fact that when there is no sub-cell correlation in those cells with δ, or when $\overline{\xi_{12}}(\delta) = 0$, the identity $[N_1 N_2] = [N_1][N_2] = [N]{\rm d}V_1/V_{\rm cell}\,[N]{\rm d}V_2/V_{\rm cell}$ is satisfied by equation (20). The net variance is therefore given as

$$\sigma^2(\delta) \equiv \left[(N - [N])^2\right] = [N] + \Delta_{\rm scc}(\delta), \quad (22)$$

which is again an application of equation 7.66 in Peebles (1993) to the conditional cases we consider. This also suggests that the true PDF deviates from the pure Poisson statistics, and the super-Poissonian PDF suggested by Saslaw & Hamilton (1984), given by

$$P_{\rm cell}(N) = \frac{[N]}{N!} {\rm e}^{-[N](1-\beta)-N\beta} (1-\beta) \left([N](1-\beta) + N\beta\right)^{N-1}, \quad (23)$$

shows excellent agreement with e.g. the distribution of N-body haloes of $M = 10^{10\text{–}11}\,{\rm M}_\odot$ (Neyrinck et al. 2014). Here $\beta \equiv 1 - \sqrt{\sigma^2(\delta)/[N]}$ represents the degree of super-Poissonianity.

Sometimes, we may only be interested in those cells that contain at least one halo. Quantifying this might be useful when haloes are rare, such that not all the cells with given δ are occupied by these haloes. It is therefore useful to have the conditional probability that there are N haloes in the cell (with δ and $V_{\rm cell}$) once a halo is found in that cell (let us denote these cells by 'active cells'). This requires renormalizing the PDF

$$\begin{aligned} P_{\rm cell}(N|N \geq 1) &\equiv P(N|\delta, V_{\rm cell};\ N \geq 1) \\ &= \frac{P_{\rm cell}(N)}{P(N \geq 1|\delta, V_{\rm cell})} \\ &= \frac{P_{\rm cell}(N)}{1 - P(N = 0|\delta, V_{\rm cell})} \\ &= \frac{P_{\rm cell}(N)}{1 - {\rm e}^{-[N](1-\beta)}}, \end{aligned} \quad (24)$$

where in the last equality we used equation (23). The mean value of N inside 'active' cells will then be given by

$$[N]_{\rm a} \equiv \sum_{N=1}^{\infty} N P_{\rm cell}(N|N \geq 1) = \frac{[N]}{1 - {\rm e}^{-[N](1-\beta)}}, \quad (25)$$

which should be used as the estimator of the mean value. Two limiting cases are noteworthy. First, when $[N] \ll 1$, $P_{\rm cell}(N|N \geq 1)$ can be approximated as

$$P_{\rm cell}(N|N \geq 1) \simeq \frac{1}{N!} {\rm e}^{-N\beta} (N\beta)^{N-1}, \quad (26)$$

which is no longer dependent on $[N]$. In the other extreme, $[N] \gg 1$, $P_{\rm cell}(N|N \geq 1) = P_{\rm cell}(N)$.

Similarly, we use the same renormalization to determine the collapsed fraction inside active cells:

$$[f_{\rm c}(\delta)]_{\rm a} = \frac{[f_{\rm c}(\delta)]}{1 - {\rm e}^{-[N](1-\beta)}}. \quad (27)$$

When $[N] \ll 1$, as the mass function is biased towards the least massive haloes, $[f_{\rm c}]_{\rm a} \simeq M_{\rm min}/M_{\rm cell} = M_{\rm min}\rho_0^{-1} V_{\rm cell}^{-1}(1+\delta)^{-1}$. When $[N] \gg 1$, $[f_{\rm c}]_{\rm a}$ trivially converges to $[f_{\rm c}]$.

Note that $[N]$ can be smaller than 1. This does not mean that we will find a fractional, less-than-unity number of haloes on average, which is simply unphysical. This means instead, assuming ergodicity, that

$$\begin{aligned} [N] &= \frac{\text{total number of haloes found in all cells with } \delta}{\text{total number of cells with } \delta} \\ &\approx \frac{\text{number of active cells with } \delta}{\text{total number of cells with } \delta}, \end{aligned} \quad (28)$$

where the approximation is made possible due to the fact that when $[N] \ll 1$, the PDF $P_{\rm cell}(N|N \geq 1)$ is peaked at $N = 1$.

4 RESULTS

4.1 Mean unconditional halo mass function

The mean, unconditional halo mass functions at both high and low redshifts were recently discussed in detail in Watson et al. (2014), much of it based on the same simulations as the current work. Therefore, we will only summarize a selection of the mean mass function properties that are most relevant here.

In Fig. 1 we show the mass functions in the mass range $M \geq 10^5\,{\rm M}_\odot$ at selected redshifts based on the $L_{\rm box} = 20$ and $6.3\,h^{-1}$ Mpc simulations, together with PS and ST analytical mass functions. The actual quantities plotted are halo number densities $\Delta n \equiv \int_{M_1}^{M_2} ({\rm d}n/{\rm d}M){\rm d}M$, integrated over equal-size logarithmic mass bins. The last mass bin includes all haloes with mass $M \geq 10^9\,{\rm M}_\odot$. The two simulated mass functions show excellent agreement with each other, except for the high-mass end, where the mass function is truncated due to finite volume. This agreement indicates the consistency of the N-body simulation over varying box size.

Compared to the analytical expressions, our N-body mass functions are in better agreement with ST than PS mass functions. The agreement with ST at all redshifts is within ∼25 per cent for $M = [10^5\text{–}10^6]\,{\rm M}_\odot$, the haloes in which range numerically dominate the minihalo population. At very high redshifts, $z \gtrsim 20$, ST mass function slightly overpredicts halo population at $M = [10^5\text{–}10^{5.5}]\,{\rm M}_\odot$ and underpredicts halo abundance at $M \geq 10^6\,{\rm M}_\odot$, with tendency to deviate increasingly as M increases, while at relatively low redshifts, overprediction occurs at $M = [10^{5.5}\text{–}10^6]\,{\rm M}_\odot$. As discussed in Watson et al. (2014) these differences are partly due to our usage of a halo finder based on spherical overdensity instead of the friends-of-friends one used by ST, and also to the limitations of the ST fit which was based on low-redshift data and relatively small simulations. In contrast, the classical PS mass function gives a poor fit to N-body minihalo data at all redshifts, severely underpredicting the abundance of rare ($\nu = \delta_{\rm c}/\sigma_M \gg 1$) haloes and overpredicting the abundance of $\nu \ll 1$ haloes. Only for the most common ($\nu \approx 1$) haloes PS is a more reasonable approximation (and also agrees with ST).

Assuming that the prescription for the conditional mass function (linking equation 12 with unconditional mass function) provides a

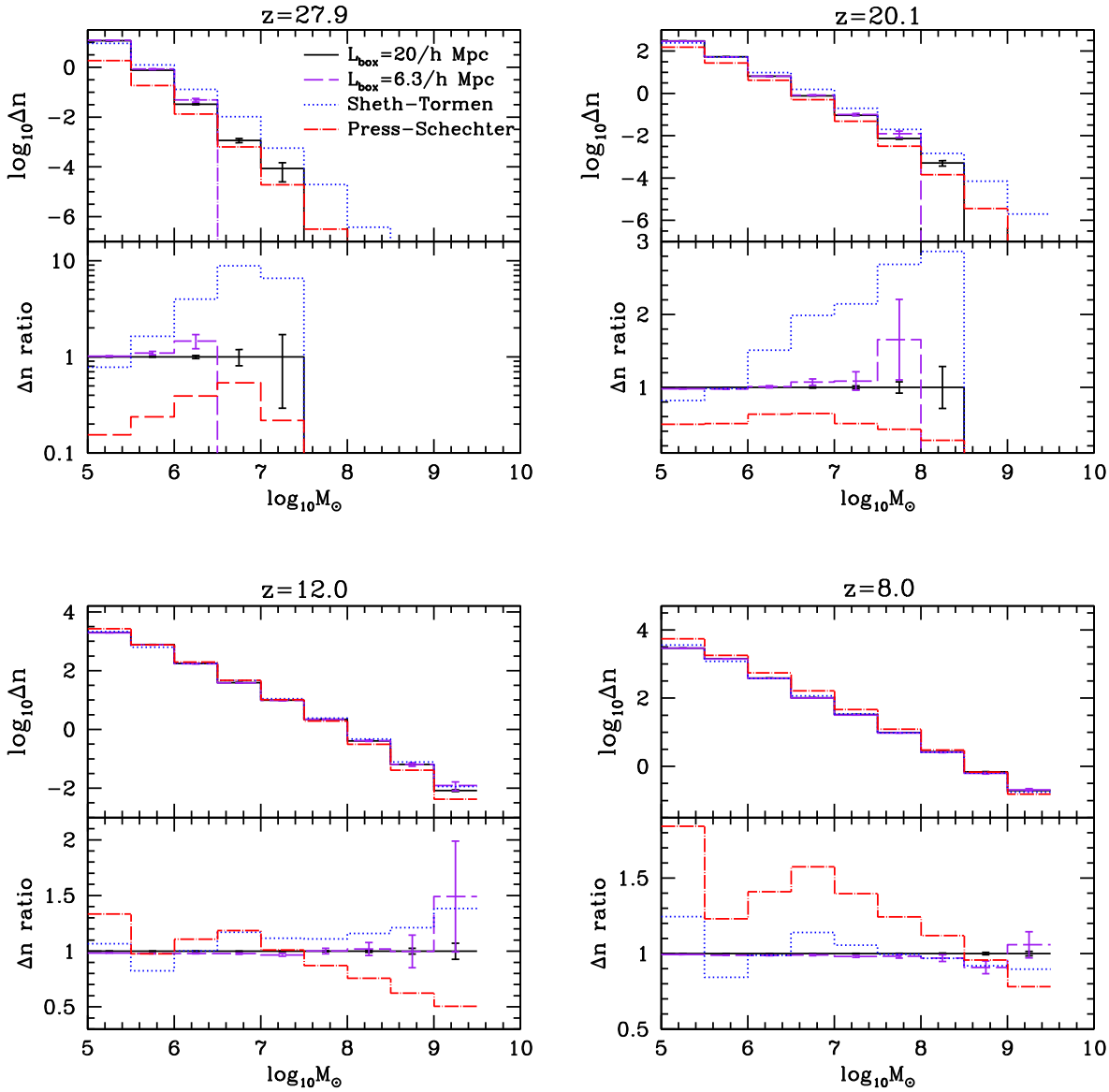

Figure 1. Unconditional, mean mass functions of minihaloes found in small-box N-body simulations. Plotted are the mass functions (each top panel) inside $20\,h^{-1}$ Mpc (black, solid) and $6\,h^{-1}$ Mpc (purple, dashed) boxes, along with the Sheth–Tormen (blue, dotted) and Press–Schechter (red, dot–dashed) mass functions, all integrated over equal-size logarithmic mass bins, and the ratios of these mass functions (each bottom panel with the same line types) to the mass function inside the $20\,h^{-1}$ Mpc box. The error bars represent 1σ standard deviation in each mass bin.

correct theoretical framework, one may expect that a good fit to unconditional mass function will also provide a good fit to conditional mass function when combined with equation (12). Therefore, we can expect that $(\mathrm{d}n/\mathrm{d}M)_{N-\mathrm{body,b}}$ will be the best fit to the mean conditional mass function from the simulations, and $(\mathrm{d}n/\mathrm{d}M)_{\mathrm{ST,b}}$ will also be a good fit, while $(\mathrm{d}n/\mathrm{d}M)_{\mathrm{PS,b}}$ will be a poor fit. We will test this expectation in Section 4.2.

4.2 Mean biased halo mass function

We now show how the mean, conditional mass functions of N-body haloes behave in terms of δ, and compare this to the modelling predictions based on the different mass functions, $(\mathrm{d}n/\mathrm{d}M)_{\mathrm{PS,b}}$, $(\mathrm{d}n/\mathrm{d}M)_{\mathrm{ST,b}}$ and $(\mathrm{d}n/\mathrm{d}M)_{N-\mathrm{body,b}}$. We also compare these to the model based on the linear bias. The stochasticity in this relation will be treated in Section 4.3.

4.2.1 Minihaloes

Minihaloes are usually defined by their hydrodynamical properties. Their minimum mass is the cosmic Jeans mass determined by the mean IGM temperature, and their maximum mass is the mass of haloes whose virial temperature is about 10^4 K. While this is the general definition, the uncertainty of the mean IGM temperature at high redshift makes the definition of the minimum mass somewhat uncertain. In this work we instead take their mass to be in a fixed range $M = [10^5\text{–}10^8]\,\mathrm{M}_\odot$, which is of more direct use to N-body data at fixed mass resolution. Both 6.3 and $20\,h^{-1}$ Mpc boxes resolve haloes down to $M = 10^5\,\mathrm{M}_\odot$. The latter simulation thus provides a better statistics by encompassing a volume 32 times as large as that of the former one.

We first examine how well the models based on the analytical mass function fits match the N-body data. Figs 2 and 3 show

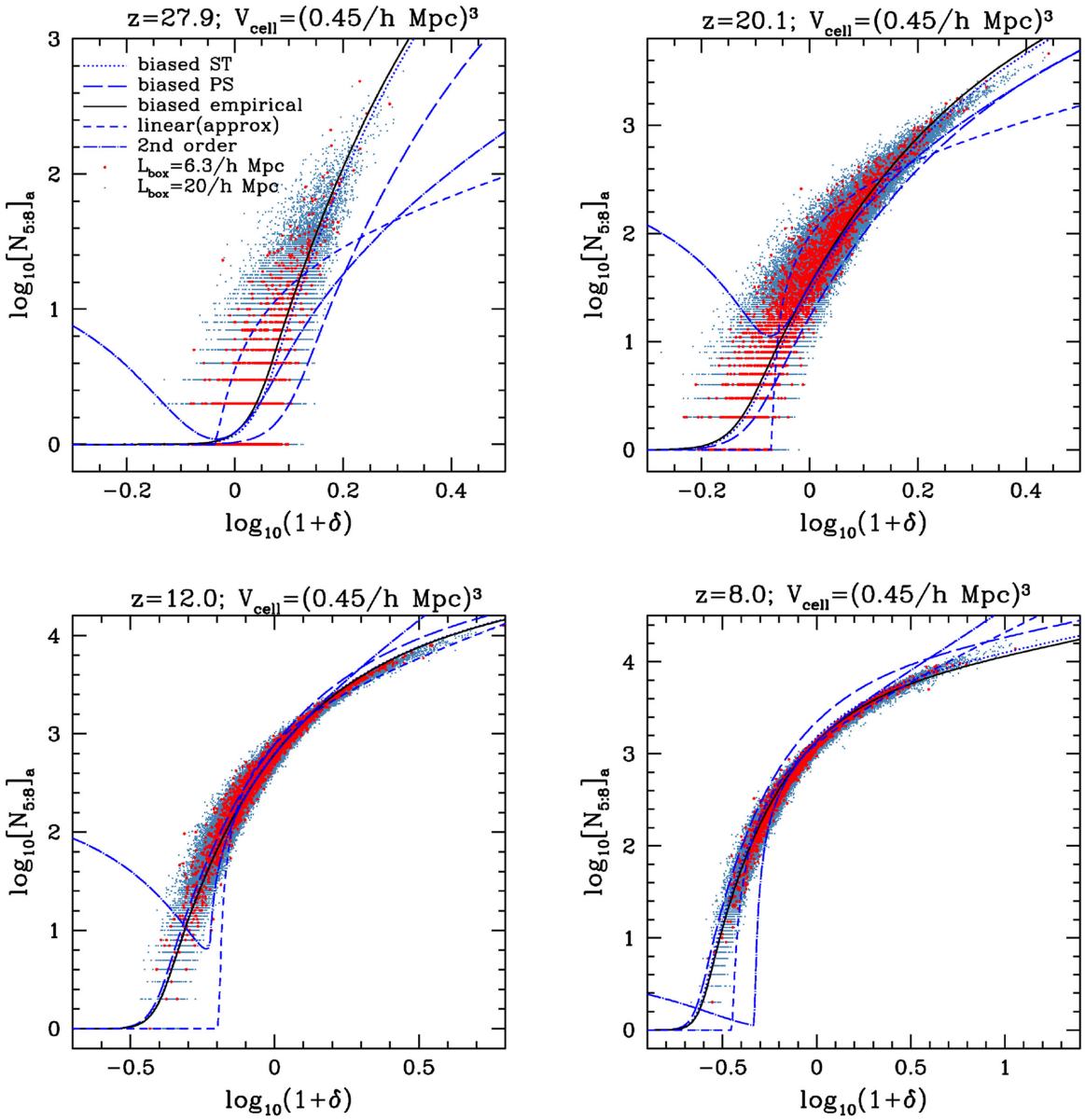

Figure 2. Correlation between the number of minihaloes $N_{5:8}$ and the cell overdensity δ, when the Eulerian volume of the cell is chosen to be $(0.45\,h^{-1}\,\mathrm{Mpc})^3$. Data points are from N-body simulations in the 6.3 h^{-1} Mpc box (red, larger dot) and 20 h^{-1} Mpc box (slate blue, smaller dot), which are sampled by 14^3 and 45^3 cells, respectively. Theoretical predictions for active cells (equation 25) based on $(\mathrm{d}n/\mathrm{d}M)_{N-\mathrm{body,b}}$ (solid, black; equation 15), $(\mathrm{d}n/\mathrm{d}M)_{\mathrm{ST,b}}$ (dotted, blue; equation 14), $(\mathrm{d}n/\mathrm{d}M)_{\mathrm{PS,b}}$ (long-dashed, blue; equation 9), the one by the linear bias approximation without the 0-point offset B_0 defined in Section 4.4 (short-dashed, blue; equation 13) combined with $(\mathrm{d}n/\mathrm{d}M)_{N-\mathrm{body,b}}$ and the one by the second-order approximation with B_0 (dot–dashed, blue; equations 30–35) also combined with $(\mathrm{d}n/\mathrm{d}M)_{N-\mathrm{body,b}}$ are plotted for comparison.

the analytical estimates and N-body data on the total number of minihaloes $[N_{5:8}]_\mathrm{a}$ under different Eulerian cell sizes. We find that the numerical data from the two simulation volumes is in excellent agreement and that$(\mathrm{d}n/\mathrm{d}M)_{\mathrm{ST,b}}$ and $(\mathrm{d}n/\mathrm{d}M)_{N-\mathrm{body,b}}$ fit the N-body data well over almost the entire range of δ and z, while $(\mathrm{d}n/\mathrm{d}M)_{\mathrm{PS,b}}$ and the linear relation $\delta_\mathrm{h} = b_\mathrm{lin}\delta$ both provide poor fits to the data in general. Even though $[N_{5:8}]_\mathrm{a}$ and $[f_{\mathrm{c},5:8}]_\mathrm{a}$ are integral quantities, given that smallest-mass haloes numerically dominate the halo population, both the data and semi-analytical estimates reflect predominantly the low-mass end. Note that as seen in Fig. 1, $(\mathrm{d}n/\mathrm{d}M)_{\mathrm{ST}}$ agrees well with $(\mathrm{d}n/\mathrm{d}M)_{N-\mathrm{body}}$ in the low-mass end, and this is the reason why $(\mathrm{d}n/\mathrm{d}M)_{\mathrm{ST,b}}$ provides a good fit. If we focused on the high-mass end only, $(\mathrm{d}n/\mathrm{d}M)_{\mathrm{ST,b}}$ would be a very poor fit to the observed bias, because the average ST mass function $(\mathrm{d}n/\mathrm{d}M)_{\mathrm{ST}}$ has large discrepancy from the actual N-body data for e.g. $M \geq 10^7\mathrm{M}_\odot$. In contrast, the collapsed fraction in haloes $[f_{\mathrm{c},5:8}]_\mathrm{a}$ (Figs 1 and 2 of the Supplementary Material) is a mass-weighted quantity and thus reflects the high-mass end better than does $[N_{5:8}]_\mathrm{a}$, but the rapid exponential cut-off of the mean halo mass function $\mathrm{d}n/\mathrm{d}M$ at increasing M still moderates the contribution from the high-mass haloes. The similarity between $(\mathrm{d}n/\mathrm{d}M)_{\mathrm{ST,b}}$ and $(\mathrm{d}n/\mathrm{d}M)_{N-\mathrm{body,b}}$ reflect the simple fact that the unconditional mass functions, $(\mathrm{d}n/\mathrm{d}M)_{\mathrm{ST}}$ and $(\mathrm{d}n/\mathrm{d}M)_{N-\mathrm{body}}$, are similar around the low-mass end.

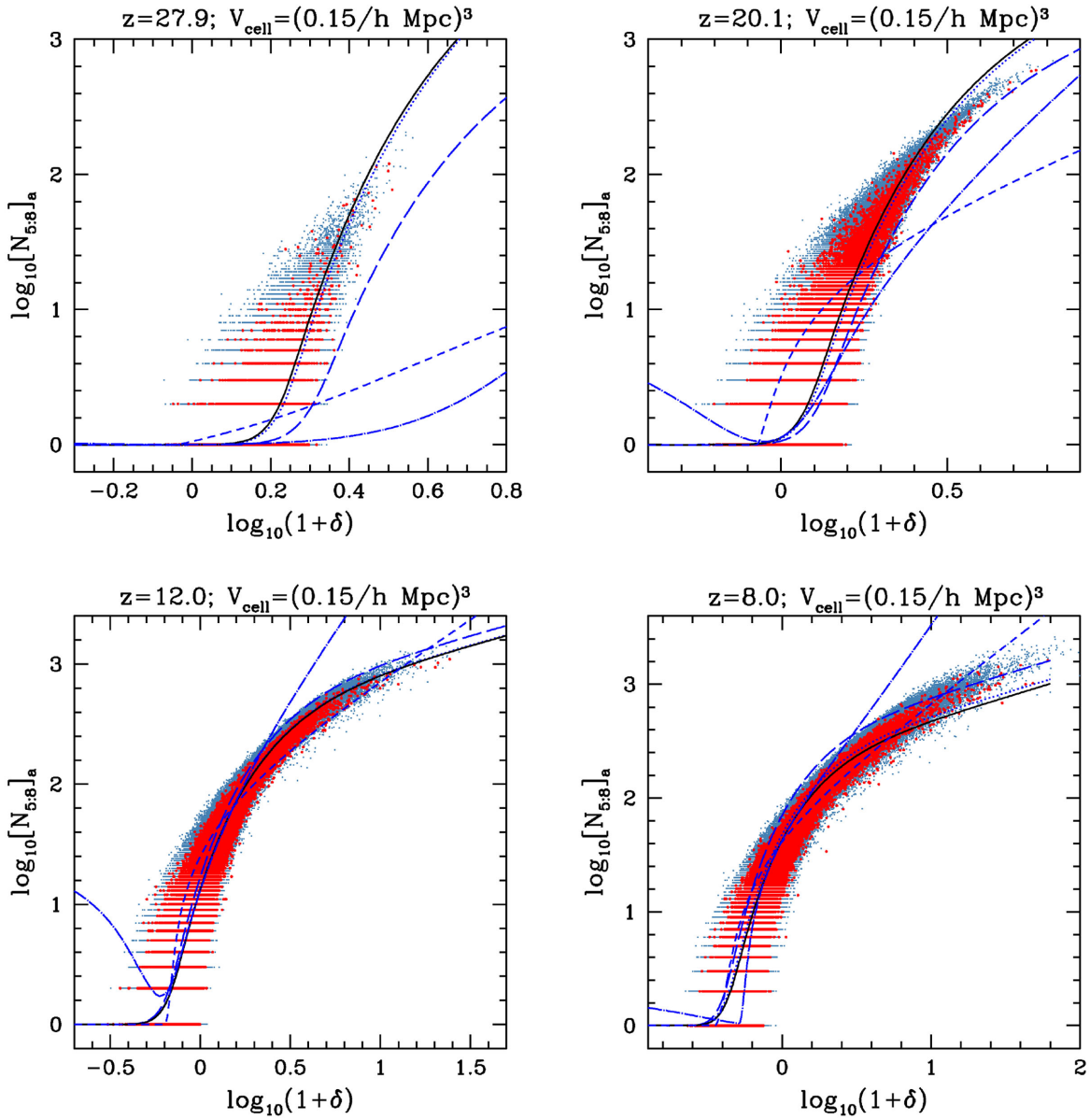

Figure 3. Same as Fig. 2, except that the volume of the cell is now $(0.15\,h^{-1}\,\mathrm{Mpc})^3$. The 6.3 h^{-1} Mpc box (red circle) and 20 h^{-1} Mpc box (green dot) are sampled by 44^3 and 135^3 cells, respectively.

As the cell size shrinks, however, some discrepancy appears at high δ regime. Both $(\mathrm{d}n/\mathrm{d}M)_{\mathrm{ST,b}}$ and $(\mathrm{d}n/\mathrm{d}M)_{N-\mathrm{body,b}}$ predictions overestimate the N-body data substantially at $\delta \gtrsim 1.5$, when the volume of the cell has shrunken from $(0.45\,h^{-1}\,\mathrm{Mpc})^3$ to $(0.15\,h^{-1}\,\mathrm{Mpc})^3$: see Fig. 3. At this point, where δ approaches the overdensity criterion for halo identification, we suspect that this could be a symptom of extreme non-linearity: the mean mass of the cell, $M_{\mathrm{cell}} = 3.8 \times 10^8\,\mathrm{M}_\odot$, is small enough to be comparable to the high-mass end of minihaloes, or $10^8\mathrm{M}_\odot$.

In summary, unless the cell is too small, and thus potentially quite non-linear, the mean non-linear bias of N-body minihaloes at high redshifts can be explained well by the simple hybrid prescriptions $(\mathrm{d}n/\mathrm{d}M)_{N-\mathrm{body,b}}$ and $(\mathrm{d}n/\mathrm{d}M)_{\mathrm{ST,b}}$. In contrast, at high redshifts, the linear relation $\delta_{\mathrm{h}} \propto \delta$ deviates too much from the N-body minihalo data to be of much practical use at least under the filtering scales of $\lesssim$ Mpc. The disagreement of $(\mathrm{d}n/\mathrm{d}M)_{\mathrm{PS,b}}$ with the N-body data is just as severe, and we expect that $(\mathrm{d}n/\mathrm{d}M)_{\mathrm{PS,b}}$ will be useless regardless of the filtering scale, because the disagreement is caused by the poorness of the mean PS mass function. It is notable that the rarity of haloes at high redshifts makes the linear relation fail even when $|\delta| \ll 1$, which will be discussed in much detail in Section 4.4.

4.2.2 *Atomically cooling haloes*

Atomically cooling haloes (ACHs hereafter) are named after the dominant cooling mechanism of baryonic gas inside. Atomic line radiation can cool primordial-composition gas to $T \simeq 10^4$ K from its initially higher virial temperature. Star formation are believed to occur inside these haloes as pre-existing metals or newly formed H_2 can further cool the gas down to $T \sim 100$ K. Therefore, ACHs are usually defined by their virial temperature: haloes with

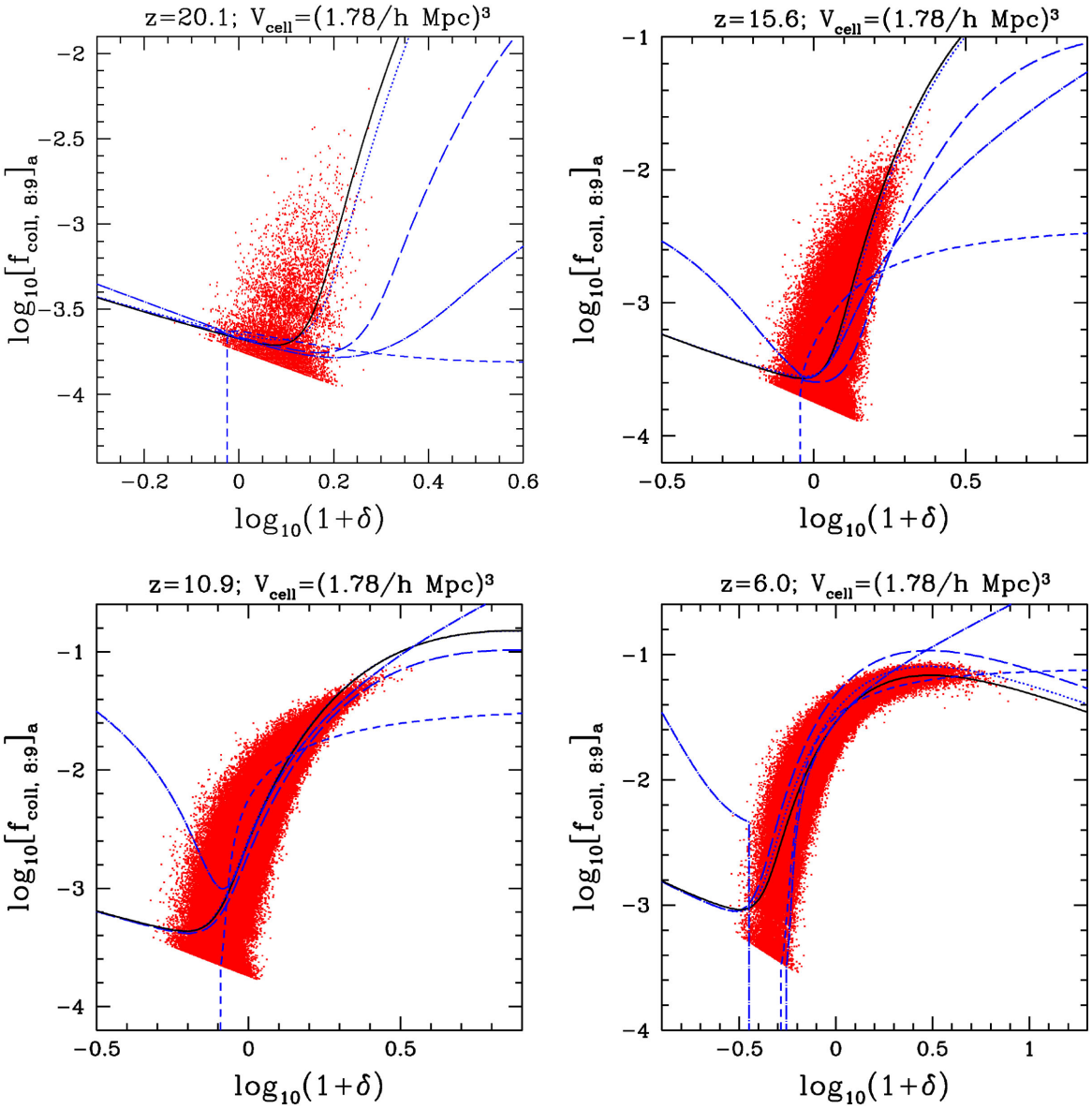

Figure 4. Correlation between the fraction of mass collapsed ($f_{\rm coll,\,8:9}$) into LMACHs ($M = 10^8$–$10^9\,{\rm M}_\odot$) and the cell overdensity δ in the $114\,h^{-1}$ Mpc box, where the box is sampled by 64^3 grid cells. Conventions for plotting follow those of Fig. 2, except for the data points (red point).

$T \gtrsim 10^4$ K. As this threshold virial temperature roughly coincides with $M \simeq 10^8\,{\rm M}_\odot$, here we define ACHs as those haloes with $M \geq 10^8\,{\rm M}_\odot$. The ACHs can be grouped further into low-mass ACHs (LMACHs), for which the gas pressure of the photoheated IGM in an ionized patch prevented the halo from capturing the gas it needs to form stars, and high-mass ACHs (HMACHs), for which gravity was strong enough to overcome this 'Jeans-mass filter' and form stars even in the ionized patches. The dividing line between LMACHs and HMACHs occurred roughly at $\sim 10^9\,{\rm M}_\odot$ (although the precise boundary value is still uncertain).

As our $114\,h^{-1}$ Mpc box simulation resolves haloes of $M \geq 10^8\,{\rm M}_\odot$, ACHs defined as above are fully identified. Even though the inner structure of low-mass end haloes is not resolved near the resolution limit (see Section 2), for our considerations only the number count of haloes matters, both for the mean halo bias and stochasticity[3] and therefore our results are not affected by this.

We choose two filtering scales, $114/h/64 = 1.78\,h^{-1}$ Mpc and $114/h/32 = 3.56\,h^{-1}$ Mpc. While these choices are somewhat arbitrary, we increased the filtering scales for ACHs from those for mini-haloes, due to the increased rarity of ACHs. The halo-collapsed fraction is plotted in Figs 4 and 5. While LMACHs have a finite range in

[3] As to be seen in Section 3.4 and Section 4.3, the conditional halo correlation function determines the stochasticity. The halo correlation function is composed of the one-halo term and the two-halo term, and the dominant contribution to stochasticity comes from the two-halo term. Therefore, it is not required to fully resolve the halo structure in estimating the stochasticity.

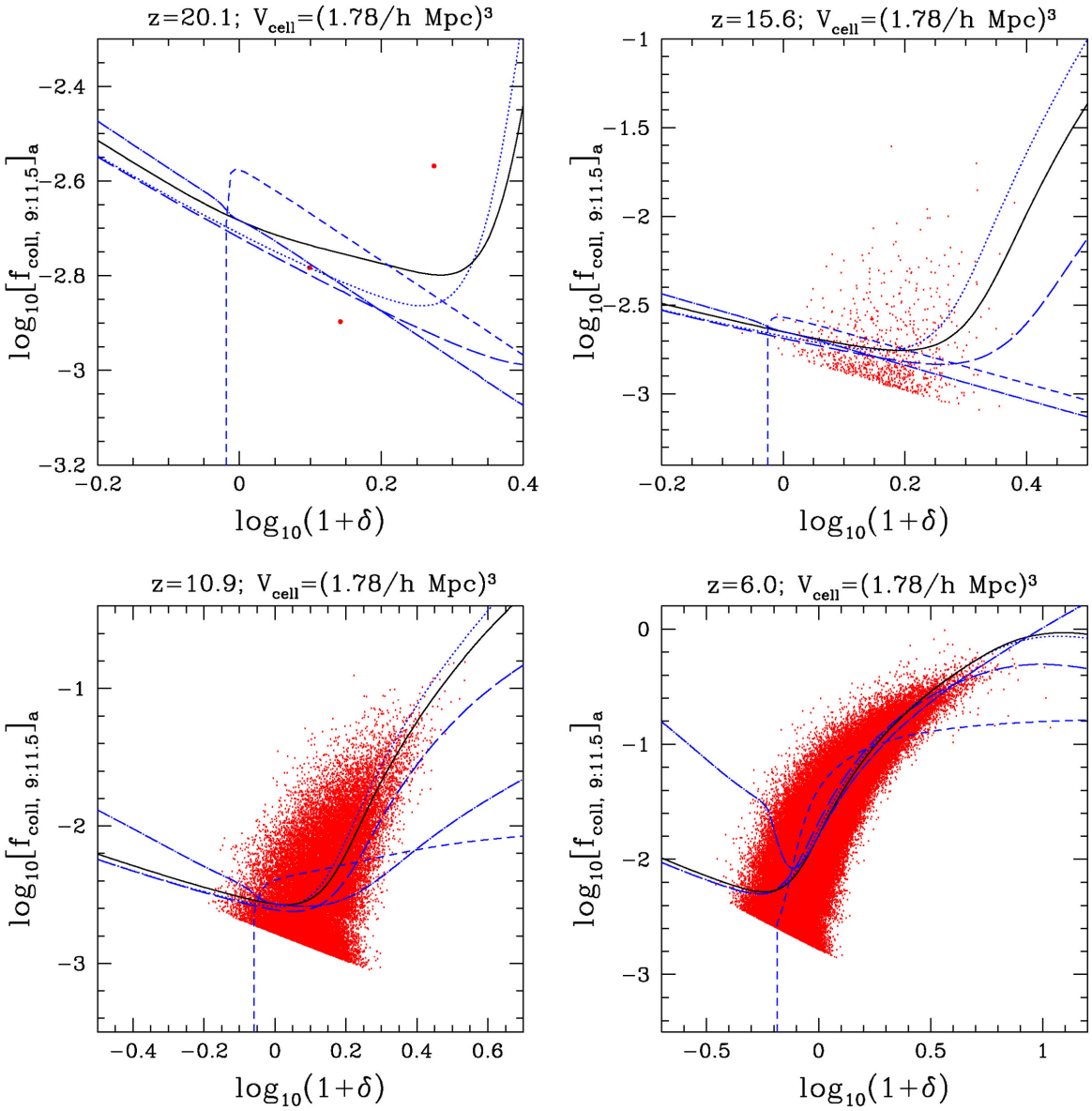

Figure 5. Correlation between the fraction of mass collapsed ($f_{\rm coll,\,9:11.5}$) into HMACHs ($M = 10^9$–$10^{11.5}\,{\rm M}_\odot$), where the maximum mass is roughly the mass of a cell, and the cell overdensity δ in the 114 h^{-1} Mpc box, where the box is sampled by 64^3 grid cells. Conventions for plotting follow those of Fig. 2, except for the data points (red point).

mass, because HMACHs are defined to have a loose end, we assign their maximum mass as the one somewhat smaller than the mass of the average-density cell: $M_{\rm max} = 10^{11.5}\,{\rm M}_\odot$ and $M_{\rm max} = 10^{12.5}\,{\rm M}_\odot$ for cells with $V_{\rm cell} = (1.78\,h^{-1}\,{\rm Mpc})^3$ and $V_{\rm cell} = (3.56\,h^{-1}\,{\rm Mpc})^3$, respectively. Otherwise, the bias formalism breaks down (equation 5). Overall, the mean values of both the LMACH collapsed fraction ($[f_{\rm c,\,8:9}]_{\rm a}$), and the HMACH collapsed fraction ($[f_{\rm c,\,9:11.5}]_{\rm a}$ and $[f_{\rm c,\,9:12.5}]_{\rm a}$) are well predicted by equation (16) when we adopt $({\rm d}n/{\rm d}M)_{N-{\rm body,b}}$ (equation 15). For LMACHs, $({\rm d}n/{\rm d}M)_{\rm ST,b}$ provides as good a fit as $({\rm d}n/{\rm d}M)_{N-{\rm body,b}}$, except at $z = 6$ where ST prescription somewhat overestimates the mean. For HMACHs, the biggest discrepancy between $({\rm d}n/{\rm d}M)_{\rm ST,b}$ and $({\rm d}n/{\rm d}M)_{N-{\rm body,b}}$ exists at higher redshifts (e.g. $z = 15.6$) at $\log(1+\delta) \gtrsim 0$: here the small number of sampled cells at high cell-density makes it difficult to conclude which prescription provides a better estimator for the mean bias. PS prescription provides a very poor fit at all redshifts.

The linear bias parameter, for both LMACHs and HMACHs, fails in predicting the mean bias in general. This is noteworthy because even in the linear regime, including the point $\delta = 0$, the linear bias parameter predicts the bias to be off from the observed values, which was also the case for minihaloes. We discuss this issue in detail in Section 4.4.

In summary, even though LMACHs and HMACHs are very rare in the regime we study, the non-linear bias prescription combined with the mean N-body halo mass function fits the observed mean halo bias very well throughout the ranges of redshift and cell density we observe. Therefore, this hybrid bias prescription can be applied for astrophysical and cosmological applications in general. We have indeed applied the bias prescription from this work in simulating cosmic reionization by ACHs in a very large box, $425\,h^{-1}$ Mpc, in order to populate Eulerian cells with size $425/h/504 = 0.843\,h^{-1}$ Mpc (Iliev et al. 2014). Because the halo mass resolution of the

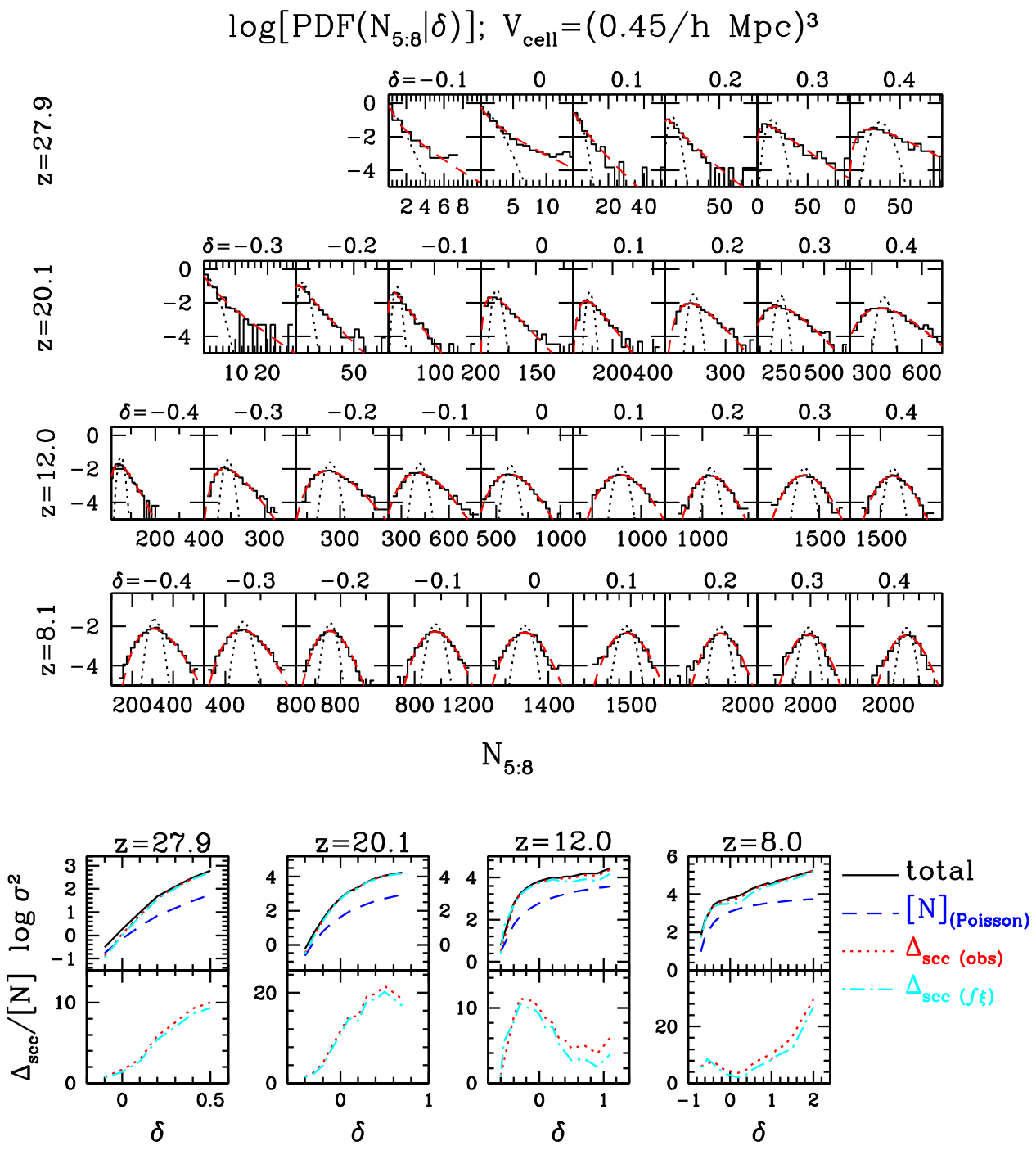

Figure 6. (a) PDFs of the number of minihaloes $N_{5:8}$ at given overdensity δ (denoted on top of each sub-panel) of cells with Eulerian volume $(0.45\,h^{-1}\,\mathrm{Mpc})^3$ observed in 20 h^{-1} Mpc box. The horizontal and vertical axes represent $N_{5:8}$ and $\log P_{\rm cell}(N_{5:8}|\delta)$, respectively. The data from simulation (histogram) is compared to the pure Poisson PDF (black, dotted) and the super-Poissonian PDF (equation 23; red, dashed). (b) Variances. Plotted are the observed total variance $(N-[N])^2$ (black, solid), the purely Poissonian variance $[N]$ (blue, dashed), the observed excess $\Delta_{\rm scc\,(obs)} = (N-[N])^2 - [N]$ (red, dotted; equation 20) and a value calculated from the sub-cell correlation function, $\Delta_{\rm scc\,(\int\xi)}$ (cyan, dot–dashed; equation 19). The range of δ and z are selected such that the number of cells with given δ (binned properly as described in Section 4.3) at given z exceed 200 for the statistical reliability of the calculated variances. The ratio $\Delta_{\rm scc}/[N]$, plotted in the bottom panels, quantifies the excess of variance over the purely Poissonian one, $[N]$.

corresponding N-body simulation was only $10^9\,\mathrm{M}_\odot$, we assigned each cell the missing LMACHs using the mean conditional mass function $(\mathrm{d}n/\mathrm{d}M)_{N-\mathrm{body,b}}$, where the LMACH mean mass function from our $114\,h^{-1}$ Mpc simulation was used to generate $(\mathrm{d}n/\mathrm{d}M)_{N-\mathrm{body,b}}$.

4.3 Stochasticity

The average behaviour of conditional mass function is well understood in terms of the biased mass function $(\mathrm{d}n/\mathrm{d}M)_{N-\mathrm{body,b}}$. Now, how does the scatter of correlation around the mean compare to the expected stochasticity? We showed in Section 3.4 that the variance

of the number of haloes N and given δ deviate from the simple Poisson value $[N]$ by the amount $\Delta_{\rm scc}(\delta)$. We now show the result of simulation and compare this to the Poisson statistics (equation 24) and $\Delta_{\rm scc}(\delta)$ (equation 19) by explicitly calculating the sub-cell-scale correlation function (equation 20).

In Fig. 6 (see also Figs 5–9 in Supplementary Material) we show the actual PDF and compare it to the expected Poisson distribution. We find that the empirical PDF does not follow pure Poisson distribution in general: the observed PDFs usually show large outliers compared to the Poisson distribution, and there is no convincing case with variance *smaller* than the Poissonian even though such a case is possible if haloes are anticorrelated under given density environment (equations 19–22). For example, Fig. 6 shows PDFs of minihalo population inside the $6.3\,h^{-1}$ Mpc box at different redshifts and δs. In order to get the distribution, each chosen δ has some width $\Delta\delta$ such that cells are chosen if their overdensity lies inside $[\delta - \Delta\delta/2, \delta + \Delta\delta/2]$. $\Delta\delta$ is taken to be narrow enough to guarantee that the PDF in each bin is a fair representation of the true PDF, while at the same time wide enough to generate a large number of cells for statistically reliable measure of the variance.

We also quantify the relative contribution of $\Delta_{\rm scc}$ to $\sigma^2(\delta)$ by the ratio $\Delta_{\rm scc}/[N]$ (note that it is compared not to $[N]_{\rm a}$ but $[N]$), in order to see the degree of deviation of PDF from the pure Poisson distribution. There are several notable features in $\Delta_{\rm scc}/[N]$. (1) At a given redshift, the ratio $\Delta_{\rm scc}/[N]$ decreases as mass of haloes increases, and thus LMACHs and HMACHs show much weaker outliers progressively. Fig. 6 and Figs 5–9 in Supplementary Material show this trend: minihaloes have $\Delta_{\rm scc}/[N] \simeq [0, 30]$, LMACHs have $\Delta_{\rm scc}/[N] \simeq [0, 6]$ and HMACHs have $\Delta_{\rm scc}/[N] \simeq [0, 2]$. (2) As one increases the filtering scale – or the size of cells – $\Delta_{\rm scc}/[N]$ tends to decrease overall. We nevertheless have some exceptions in this trend for minihaloes at very high density cells ($\delta \simeq 10$). (3) $\Delta_{\rm scc}/[N]$ is not a monotonically increasing or decreasing function of δ. (4) $\Delta_{\rm scc}$ is mostly positive both in underdense and overdense cells, indicating that the sub-cell correlation is overall positive in both regimes (see equation 20; this does not mean that there are no negative values in $\overline{\xi_{12}}(\delta)$). This contradicts the claim by Somerville et al. (2001), where they usually find that $\Delta_{\rm scc} < 0$ in underdense regions and $\Delta_{\rm scc} > 0$ in positive regions which led them to conclude that the correlation function is negative inside underdense regions and positive in overdense regions. We believe that this discrepancy comes from the erroneous definition of $\Delta_{\rm scc}$ in Somerville et al. (2001), where they subtracted the global mean number of haloes $\langle N\rangle$ averaged over all cells of δ such that $\Delta_{\rm scc} = \sigma^2(\delta) - \langle N\rangle$, while one should indeed define this as in equation (22) to reflect the effect of the sub-cell correlation function. The observed anticorrelation of $\overline{\xi_{12}}(\delta)$, or negative values of $\overline{\xi_{12}}(\delta)$ when r becomes comparable to the cell size as seen in Fig. 7 (see also Figs 10–14 in Supplementary Material), is due to the finite cell size, because any correlation existing inside a cell should be counterbalanced by anticorrelation in order to conserve the halo number.

We explicitly calculate $\overline{\xi_{12}}(\delta)$ defined by equation (20) and $\Delta_{\rm scc}(\delta)$ from equation (19). Towards this, we place a uniform grid with 25^3 sub-cells on each cell with δ, such that $\mathrm{d}V_1 = \mathrm{d}V_2 = V_{\rm cell}/25^3$. We then sample all sub-cell pairs with given distance r_{12} – discretized as the distance between centres of sub-cells – and calculate $\overline{\xi_{12}}(\delta)$ using equation (20) and $\Delta_{\rm scc}(\delta)$ using equation (19). We compare this value to the observed, residual variance $\Delta_{\rm scc} = \sigma^2(\delta) - [N]$, which are shown in the bottom panels of Fig. 6, denoted by $\Delta_{\rm scc\,(\int\xi)}$ and $\Delta_{\rm scc\,(obs)}$, respectively. The agreement between the two quantities are excellent, and thus proves the fact that $\overline{\xi_{12}}(\delta)$ is the sole origin for the super-Poissonian (or sometimes sub-Poissonian) variance in $N(\delta)$ (see also Figs 5–9 in Supplementary Material). Due to halo-number conservation, the correlation function is composed of positive (correlation) and negative (anticorrelation) parts as seen in Fig. 7 (and Figs 10–14 in Supplementary Material).

4.4 Bias in Perturbative Schemes

Local halo bias is often calculated or fitted in perturbative way, i.e. as a polynomial series of δ:

$$\delta_{\rm h} = \sum_{n=0}^{\infty} \frac{b^{(n)}}{n!}\delta^n, \tag{29}$$

where the bias parameter $b^{(n)}$ is now defined as an nth-order moment in this expansion. In practice, one should truncate the series by limiting $\delta < 1$ such that higher order moments decay more rapidly than a few lowest-order moments. In this section, we revisit the perturbative scheme by MW and examine $b^{(n)}$ in more detail.

Linear bias approximation, $\delta_{\rm h} \propto \delta$, is widely used in literature and in practical applications such as galaxy surveys for cosmology. Here, the linear bias parameter $b_{\rm lin}$ is useful when the mass of haloes is fixed, because then $b_{\rm lin}$ is a simple constant coefficient for varying δ, or $\delta_{\rm h} = b_{\rm lin}\delta$, and the same relation applies to k-space bias such that $\delta_{\rm h}(\boldsymbol{k}) = b_{\rm lin}\delta(\boldsymbol{k})$. Its limitation, however, has already been pointed out by MW themselves, by expanding the non-linear relation (equation 12) to second order in δ and first order in $\sigma^2_{\rm cell}/\sigma^2_M$. Such expansion (and truncation at some order) is useful in observing the halo bias in k-space because algebraic connection between real-space parameters and k-space parameters is possible, and also in understanding the generic behaviour of non-linear bias. We therefore examine the Taylor-expanded form of equation (12). The main difference from MW is that we expand the non-linear relation to second order in δ but keeping $\sigma^2_{\rm cell}/\sigma^2_M$-dependence accurate, because we sometimes reach $\sigma^2_{\rm cell}/\sigma^2_M \lesssim 1$. This will enable us to examine the dependence of non-linear bias on the filtering scale more accurately.

We thus Taylor-expand $\delta_{\rm h}(M|\delta)$ to the second order in δ while keeping the dependency on $R_{\rm cell}$ accurate (as in equation 30 of MW):

$$\delta_{\rm h}(M|\delta) = B_0 + B_1\delta + \frac{1}{2}B_2\delta^2, \tag{30}$$

where we use $\delta_{\rm lin} = \delta + c\delta^2$ as an expansion of $\delta_{\rm lin}$ ($c = -0.805$; see MW) and use the chain rule $(\partial/\partial\delta) = (1 + 2c\delta)(\partial/\partial\delta_{\rm lin})$. Using equations (5) and (12), we obtain[4]

$$B_0 = p^{-\frac{3}{2}}\mathrm{e}^{-q} - 1, \tag{31}$$

$$B_1 = p^{-\frac{3}{2}}\mathrm{e}^{-q}\left(1 + \frac{p^{-1}\nu^2 - 1}{\delta_{\rm c}}\right) \tag{32}$$

[4] Rigorously speaking, in this derivation, we assume that the filtering scale $R_{\rm cell,\,L}$ is fixed, and thus so is $\sigma^2_{\rm cell}$. Because $(1+\delta)V_{\rm cell} = \frac{4\pi}{3}R^3_{\rm cell}$, this means that the $V_{\rm cell}$ changes as $V_{\rm cell} \propto (1+\delta)^{-1}$, which is not compatible with the notion of uniform grid. If we were to apply the expanded form on uniform-grid cases instead, $V_{\rm cell}$ is fixed and thus $R_{\rm cell}$ and $\sigma^2_{\rm cell}$ change as δ changes. Additional terms due to non-vanishing $\left(\partial\sigma^2_{\rm cell}/\partial\delta\right)_{\delta=0}$ will appear on B_1 and B_2 in this case. Nevertheless, $\sigma^2_{\rm cell}$ is a very slowly varying function in δ at $|\delta| \ll 1$, and thus we expect it to be higher-order correction in δ, and simply assume that $\left(\partial\sigma^2_{\rm cell}/\partial\delta\right)_{\delta=0} = 0$ in the expansion in general.

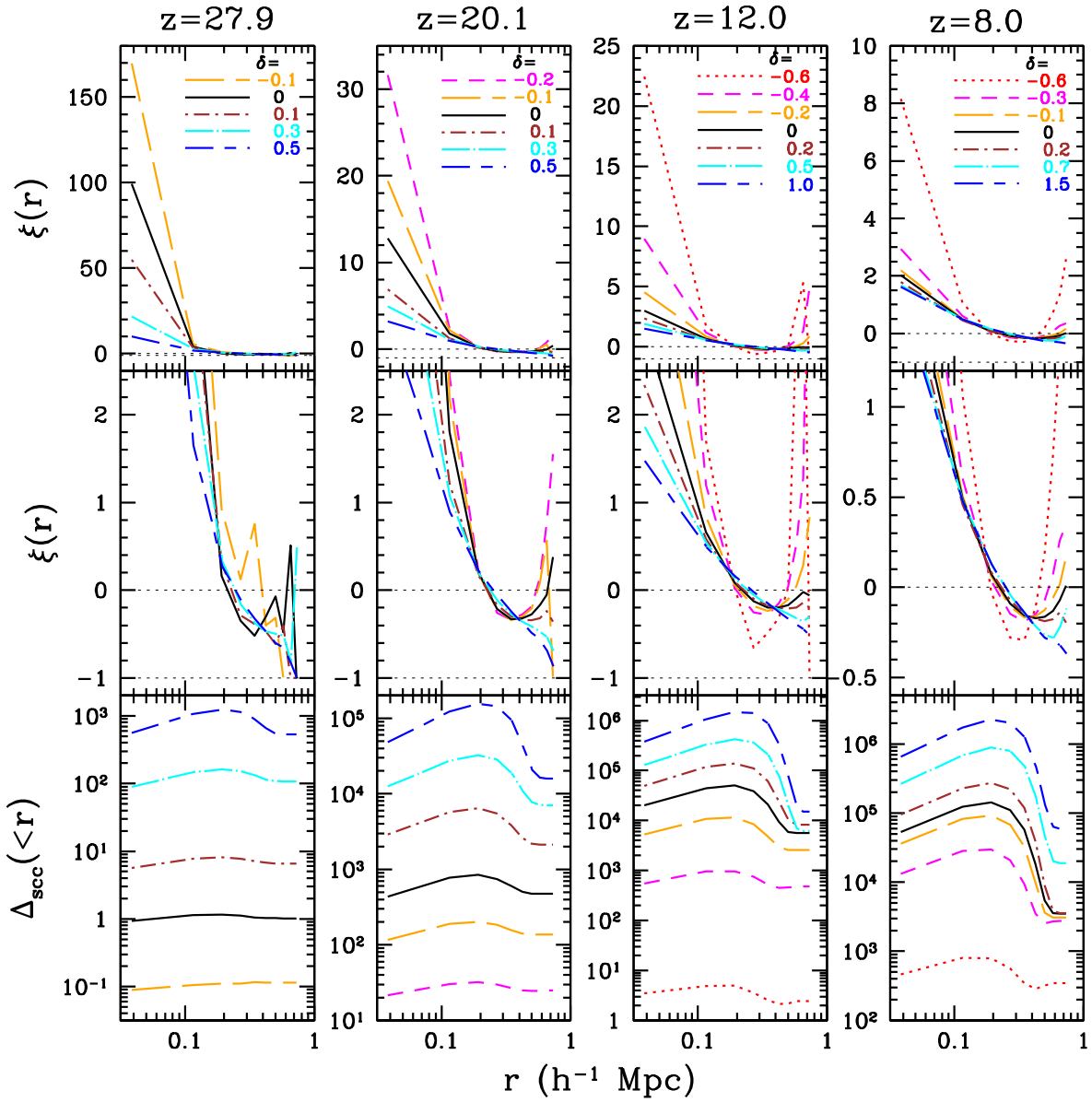

Figure 7. Conditional sub-scale correlation function $\overline{\xi_{12}}(\delta)$ of minihaloes and the cumulative contribution to $\Delta_{\rm scc}(\delta)$, inside $20\,h^{-1}$ Mpc box with $V_{\rm cell} = (0.45\,h^{-1}\,{\rm Mpc})^3$. The top- and middle-row sub-panels show $\overline{\xi_{12}}(\delta)$ as a function of two-point distance $r \equiv r_{12}$, and the bottom-row sub-panels show $\Delta_{\rm scc}(\delta; < r) \equiv \left(\frac{[N]}{V_{\rm cell}}\right)^2 \int^{<r} {\rm d}V_1 {\rm d}V_2\, \overline{\xi_{12}}(\delta; r_{12})$.

and

$$B_2 = p^{-\frac{3}{2}} {\rm e}^{-q} \left\{ \frac{p^{-1}\nu^2}{\delta_{\rm c}^2} \left(p^{-1}\nu^2 - 3\right) + \frac{2}{\delta_{\rm c}} \left(p^{-1}\nu^2 - 1\right)(1+c) \right\}, \tag{33}$$

where

$$p \equiv 1 - \frac{\sigma^2_{\rm cell,L}}{\sigma^2_{M,{\rm L}}} \tag{34}$$

and

$$q \equiv \frac{\nu^2}{2}\left(p^{-1} - 1\right). \tag{35}$$

MW approximate the dependence on $\sigma^2_{\rm cell,L}/\sigma^2_{M,{\rm L}}$ to first order, and have $B_0 = (\sigma^2_{\rm cell,L}/2\sigma^2_{M,{\rm L}})\left(3 - \nu^2\right)$, $B_1 = b_{\rm lin} = 1 + (\nu^2 - 1)/\delta_{\rm c}$ and $B_2 = (\nu^2/\delta_{\rm c}^2)\left(\nu^2 - 3\right) + (2/\delta_{\rm c})\left(\nu^2 - 1\right)(1+c)$, to which equations (31), (32) and (33) converge, respectively, when $\sigma^2_{\rm cell,L}/\sigma^2_{M,{\rm L}} \ll 1$.

B_0 explains the non-zero offsets $({\rm d}n/{\rm d}M)_{\rm b}(\delta = 0) - \langle {\rm d}n/{\rm d}M\rangle$ and $f_{\rm coll,\,b}(\delta = 0) - \langle f_{\rm coll}\rangle$ observed in almost all cases (see Figs 2–5): let us call this the '0-point offset' as MW did. 0-point offset is a natural consequence of the fact that the global mean of a quantity A, $\langle A\rangle$, differs from the selective average, $[A]_{\delta=0}$, only over cells with $\delta = 0$. If one is to apply a simple linear relation $\delta_{\rm h} \propto \delta$, it is presumed that $({\rm d}n/{\rm d}M)_{\rm b}(\delta = 0) = \langle {\rm d}n/{\rm d}M\rangle$ (or $f_{\rm coll,\,b}(\delta = 0) = \langle f_{\rm coll}\rangle$) because $\delta_{\rm h}(\delta) = b_{\rm lin}\delta$ with $b_{\rm lin}$ as a constant coefficient. However, even in the linear regime in general, $\delta_{\rm h}(\delta) = B_0 + B_1\delta$ with non-zero B_0. B_0 depends strongly on ν. The negative sign of B_0 reflects the fact that the rarer the haloes, or the higher the ν, the smaller the chances are to find them in the mean-density environment ($B_0 < 0$); in the opposite

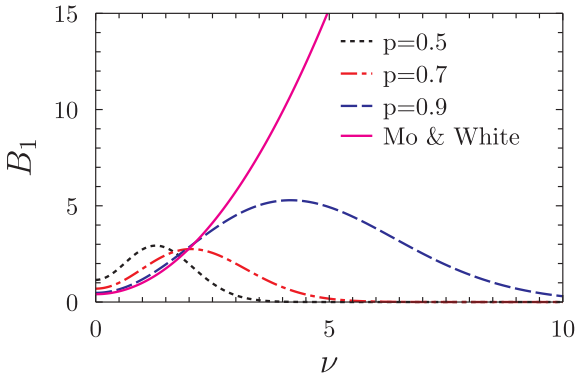

Figure 8. Linear bias parameter B_1 as a function of ν and p (equation 32). The linear bias parameter of MW corresponds to a case with $p = 1$ ($q = 0$ accordingly). Note that B_1 is in general a non-monotonic function of ν: both very abundant ($\nu \ll 1$) and very rare ($\nu \gg 1$) haloes are weakly biased to the first order in δ, while the commonly used $b_{\rm lin} = B_1(p = 1)$ of MW is a monotonic function of ν. Because $p \neq 1$ in practice, care needs to be taken when using $b_{\rm lin}$ for very rare haloes.

regime when ν is small, $B_0 > 0$, which means that haloes are more abundant in the mean-density cells than the mean value. The sign of B_0 also indicates, under a given filtering scale, the 'overall' tendency of halo distribution: when $B_0 < 0$, the net number of haloes found in overdense regions is larger than that in underdense regions, and when $B_0 > 0$, the net number of haloes found in overdense regions is smaller than that in underdense regions. As a practical example, it will be very important to study haloes in voids if those haloes are a very abundant type, or $\nu \ll 1$.

It is important to note that if the bias function (equation 10) is expanded instead of $\delta_{\rm h}$, one should include the singular term B_0/δ such that $b = B_0/\delta + B_1$, because otherwise the approximated linear bias parameter cannot explain the offset. In this sense, b should not be taken as a physical quantity but merely as a mathematical entity representing the fully non-linear dependence of $\delta_{\rm h}$ on δ. $\delta_{\rm h}$ ($=b\delta$) is a physical quantity which does not become singular when $\delta \to 0$.

B_1 is a good indicator of the overall trend of bias. The sign of B_1, which is always positive, guarantees that haloes are not anti-biased but biased for higher cell-densities regardless of ν or the filtering scale, as long as δ is in the linear regime. B_1 depends on both ν and $\sigma^2_{R_{\rm cell,m}}/\sigma^2_M$. At fixed halo mass and filtering scale, B_1 increases as ν increases when $\nu < \nu_{\rm crit,lin} \equiv p\sqrt{1-\delta_{\rm c}+2/(1-p)}$ and decreases when $\nu > \nu_{\rm crit,\,lin}$ (Fig. 8). Such non-monotonic trend in B_1 would not be observed when filtering scale is large enough, because then $\nu_{\rm crit,\,lin} \to \infty$. At fixed halo mass (and thus fixed ν and $\sigma^2_{R_{\rm f}}$ at some z), the effect of filtering scale or p on B_1 is also a mixed bag depending on rarity of haloes (or ν) as seen in Fig. 8.

5 SUMMARY AND DISCUSSION

We investigated the local bias of cosmological halo formation in the fully non-linear regime, using both halo data from N-body simulations sampled on uniform grids and theoretical estimates for Eulerian halo bias. Over the wide dynamic range of halo mass, from 10^5 to $\sim 10^{12}\,{\rm M}_\odot$, we find that the observed biased population of haloes $({\rm d}n/{\rm d}M)_{\rm b}$ inside a cell with density δ can be matched well by the convolution of the mean N-body mass function $\langle {\rm d}n/{\rm d}M\rangle_{N-{\rm body}}$ with the non-linear bias parameter derived from the extended Press–Schechter formalism. Convolution with the PS mass function provides very poor fits in general, and convolution with the ST mass function provides fits slightly poorer than $\langle {\rm d}n/{\rm d}M\rangle_{N-{\rm body}}$. Nevertheless, as the ST mass function is known to break down for very rare haloes (see e.g. the large discrepancy of the ST mass function for haloes of $M \geq 10^6\,{\rm M}_\odot$ at $z \geq 20$ in Fig. 1), it is best to avoid both PS and ST, and instead use $\langle {\rm d}n/{\rm d}M\rangle_{N-{\rm body}}$ in convolving the mean mass function to the bias factor given by equation (12). Based on the fact that the observed bias in halo population is well matched by the hybrid estimate $({\rm d}n/{\rm d}M)_{N-{\rm body,b}}$ which combines two physical quantities with different origins (the average mass function $\langle {\rm d}n/{\rm d}M\rangle_{N-{\rm body}}$ is determined by a specific halo-identification scheme and the non-linear bias parameter is based on the extended Press–Schechter theory), this prescription should be applicable in general to cases under other halo-identification schemes.

We also find that the variance of halo numbers inside grid cells with given overdensity is not purely Poissonian, but has additional variance. This variance originates from the sub-cell-scale halo–halo correlation, which we proved quantitatively by explicitly calculating the conditional correlation functions. In the regime we studied ($z \gtrsim 6$ and unigrid filtering with cell size of $\sim$[0.2–3.6]$\,h^{-1}\,$Mpc), we find that the additional variance is always positive except for some negative values sporadically observed for haloes with $M > 10^9\,{\rm M}_\odot$.

The non-linear bias prescription described in our paper can be used to generate mock halo catalogues in the following sequence:

(i) Generate or adopt a mean mass function of haloes $({\rm d}n/{\rm d}M)_{N-{\rm body}}$. It is advised not to use the PS mass function, due to the large discrepancy from the usual N-body halo catalogues practically over the full mass range.

(ii) Generate a density field at a redshift of interest: if N-body data is available, adopt a proper smoothing scheme to generate a density field from the distribution of particles. Depending on the size of cells, cell density can become non-linear, and therefore N-body simulation is recommended.

(iii) Place a uniform grid on the density field from step (ii), and identify the comoving volume of the cell as $V_{\rm cell}$.

(iv) Visit a cell, and identify the cell overdensity δ. Use equation (4) to deduce $R_{\rm cell}$. Take $R_{\rm cell}$ as the spatial filtering scale of the linearly extrapolated density field to $z = 0$, and calculate the corresponding variance $\sigma^2_{R_{\rm cell}}$. Use equations (7) and (8) (or the numerical fit given by equation 18 of MW) to find matching $\delta_{\rm lin}$ of δ.

(v) To populate a cell with a halo of mass M, use equation (3) to obtain $R_{\rm f}$, and take this as the filtering scale of the linearized density field at $z = 0$ and calculate the corresponding variance σ^2_M.

(vi) Plug quantities from steps (iv) and (v) in equation (5), then use equation (6), then finally use equation (15) to calculate the biased halo mass function $({\rm d}n/{\rm d}M)_{N-{\rm body,b}}$. Multiplying the infinitesimal mass bin ${\rm d}M$ and $V_{\rm cell}$ to $({\rm d}n/{\rm d}M)_{N-{\rm body,b}}$, one obtains the mean number of haloes $[N]$ of $M = [M, M + {\rm d}M]$ in the cell.

(vii) Iterate steps (iv)–(vi) over all cells in the box.

(viii) If one wants to implement stochasticity, which should indeed affect the power spectrum of halo density field, use equation (23) with $[N]$ from step (vi) and an empirically found $\sigma^2(\delta)$ to include super-Poisson stochasticity and sample haloes by the Monte Carlo method. For a selected range of halo masses and cell sizes as described in Sections 4.2 and 4.3, a reader may contact us for these values.

Perturbative approach to the non-linear bias is found limited. First, one needs to be careful when approximating the halo bias by a simple linear relation $\delta_{\rm h} \propto \delta$, because even when the filtered density field is in the linear regime, $|\delta| \ll 1$, the 0-point offset (equation 31) may not be negligible. In such cases, one should of

course take B_0 into account such that $\delta \simeq B_0 + b_{\rm lin}\delta$. This 0-point offset $((\mathrm{d}n/\mathrm{d}M)_{\rm b}(\delta = 0) \neq \langle \mathrm{d}n/\mathrm{d}M \rangle)$ occurs in general when (1) haloes are rare and/or (2) the cell size is small, which MW has already recognized and we have confirmed from our data. In the non-linear regime, even the second-order perturbation, which we calculated without the approximation taken by MW (equations 31–35), provides a very poor fit in general. We thus claim that the local non-linear bias scheme should be used unless perturbative approach is unavoidable.

Non-linear bias schemes such as the one studied in this paper can be applied to both cosmological and astrophysical problems. For example, we already used the mean bias prescription in this paper as a sub-grid treatment to populate simulation boxes with haloes which are not resolved otherwise, for simulating cosmic reionization process: see Ahn et al. (2012) for populating $114\,h^{-1}$ Mpc box with minihaloes, and Iliev et al. (2014) for populating $425\,h^{-1}$ Mpc box with LMACHs. Similar approach has been attempted by de la Torre & Peacock (2013), where they test their bias-based sub-grid treatment against resolved N-body haloes in terms of two-point statistics. Their bias prescription, however, is heuristic and thus the corresponding fitting parameters should be re-evaluated when e.g. a very different dynamic range of halo mass is targeted. In contrast, even though we have just studied cosmological haloes at $z \gtrsim 6$, the agreement between data and theoretical prediction in such wide range of halo mass, cell size, cell density and redshift suggests that this prescription is valid in general.

The non-linear bias scheme studied here is valid when the primordial density field is Gaussian, and thus may not be directly used to study non-Gaussianity. It is also preferred that further study of the super-Poissonian (or sometimes sub-Poissonian) stochasticity, which we quantified here with $20\,h^{-1}$ Mpc box for minihaloes and $114\,h^{-1}$ Mpc box for LMACHs and HMACHs, is devised with higher resolution, larger-box simulations to increase statistical reliability. Stochasticity is likely to have temporal correlation as well as spatial correlation, which should be further studied for a more self-contained bias prescription.

ACKNOWLEDGEMENTS

This work was supported by a research grant from Chosun University (2010). All simulations in this work were undertaken at the Texas Advanced Computing Center (TACC) at The University of Texas at Austin under TeraGrid allocations.

REFERENCES

Adshead P., Baxter E. J., Dodelson S., Lidz A., 2012, Phys. Rev. D, 86, 063526
Ahn K., Iliev I. T., Shapiro P. R., Mellema G., Koda J., Mao Y., 2012, ApJ, 756, L16
Alvarez M. A., Busha M., Abel T., Wechsler R. H., 2009, ApJ, 703, L167
Baldauf T., Seljak U., Smith R. E., Hamaus N., Desjacques V., 2013, Phys. Rev. D, 88, 083507
Bardeen J. M., Bond J. R., Kaiser N., Szalay A. S., 1986, ApJ, 304, 15
Barkana R., Loeb A., 2004, ApJ, 609, 474 (BL)
Bond J. R., Cole S., Efstathiou G., Kaiser N., 1991, ApJ, 379, 440
Ciardi B., Ferrara A., 2005, Space Sci. Rev., 116, 625
Cole S., Kaiser N., 1989, MNRAS, 237, 1127
D'Aloisio A., Zhang J., Jeong D., Shapiro P. R., 2013, MNRAS, 428, 2765
Dalal N., Doré O., Huterer D., Shirokov A., 2008, Phys. Rev. D, 77, 123514
Datta K. K., Friedrich M. M., Mellema G., Iliev I. T., Shapiro P. R., 2012, MNRAS, 424, 762
de la Torre S., Peacock J. A., 2013, MNRAS, 435, 743
Dekel A., Lahav O., 1999, ApJ, 520, 24 (DL)
Friedrich M. M., Mellema G., Alvarez M. A., Shapiro P. R., Iliev I. T., 2011, MNRAS, 413, 1353
Furlanetto S. R., Oh S. P., 2005, MNRAS, 363, 1031
Furlanetto S. R., Zaldarriaga M., Hernquist L., 2004, ApJ, 613, 1
Harnois-Déraps J., Pen U.-L., Iliev I. T., Merz H., Emberson J. D., Desjacques V., 2013, MNRAS, 436, 540
Iliev I. T., Mellema G., Ahn K., Shapiro P. R., Mao Y., Pen U.-L., 2014, MNRAS, 439, 725
Jenkins A., Frenk C. S., White S. D. M., Colberg J. M., Cole S., Evrard A. E., Couchman H. M. P., Yoshida N., 2001, MNRAS, 321, 372
Joudaki S., Doré O., Ferramacho L., Kaplinghat M., Santos M. G., 2011, Phys. Rev. Lett., 107, 131304
Kaiser N., 1984, ApJ, 284, L9
Kitaura F.-S., Yepes G., Prada F., 2014, MNRAS, 439, L21
Lim S., Lee J., 2013, J. Cosmol. Astropart. Phys., 1, 19
Lukić Z., Heitmann K., Habib S., Bashinsky S., Ricker P. M., 2007, ApJ, 671, 1160
Manera M., Sheth R. K., Scoccimarro R., 2010, MNRAS, 402, 589
Manera M. et al., 2013, MNRAS, 428, 1036
Mao Y., D'Aloisio A., Zhang J., Shapiro P. R., 2013, Phys. Rev. D, 88, 081303
Mesinger A., Furlanetto S., Cen R., 2011, MNRAS, 411, 955
Mo H. J., White S. D. M., 1996, MNRAS, 282, 347 (MW)
Monaco P., Theuns T., Taffoni G., Governato F., Quinn T., Stadel J., 2002, ApJ, 564, 8
Monaco P., Sefusatti E., Borgani S., Crocce M., Fosalba P., Sheth R. K., Theuns T., 2013, MNRAS, 433, 2389
Neyrinck M. C., Aragón-Calvo M. A., Jeong D., Wang X., 2014, MNRAS, 441, 646
Park H., Shapiro P. R., Komatsu E., Iliev I. T., Ahn K., Mellema G., 2013, ApJ, 769, 93
Peebles P. J. E., 1993, Principles of Physical Cosmology. Princeton Univ. Press, Princeton, NJ
Press W. H., Schechter P., 1974, ApJ, 187, 425 (PS)
Reed D. S., Bower R., Frenk C. S., Jenkins A., Theuns T., 2007, MNRAS, 374, 2
Santos M. G., Amblard A., Pritchard J., Trac H., Cen R., Cooray A., 2008, ApJ, 689, 1
Saslaw W. C., Hamilton A. J. S., 1984, ApJ, 276, 13
Scoccimarro R., Sheth R. K., 2002, MNRAS, 329, 629
Shapiro P. R., Mao Y., Iliev I. T., Mellema G., Datta K. K., Ahn K., Koda J., 2013, Phys. Rev. Lett., 110, 151301
Sheth R. K., 1995, MNRAS, 274, 213
Sheth R. K., Tormen G., 1999, MNRAS, 308, 119 (ST)
Sheth R. K., Tormen G., 2002, MNRAS, 329, 61
Somerville R. S., Lemson G., Sigad Y., Dekel A., Kauffmann G., White S. D. M., 2001, MNRAS, 320, 289
Warren M. S., Abazajian K., Holz D. E., Teodoro L., 2006, ApJ, 646, 881
Watson W. A., Iliev I. T., Diego J. M., Gottlöber S., Knebe A., Martínez-González E., Yepes G., 2014, MNRAS, 437, 3776
Zahn O., Lidz A., McQuinn M., Dutta S., Hernquist L., Zaldarriaga M., Furlanetto S. R., 2007, ApJ, 654, 12

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Figure S1. Correlation between the fraction of mass collapsed into minihaloes ($f_{\rm coll}$, 5:8) and the cell overdensity δ in the 6.3/h Mpc box, where the box is sampled by 14^3 grid-cells. Conventions for plotting follow those of Fig. 2 of the main paper.

Figure S2. Correlation between the fraction of mass collapsed into minihaloes ($f_{\rm coll}$, 5:8) and the cell overdensity δ in the 6.3/h Mpc box and 20/h Mpc box, sampled by 44^3 and 135^3 cells, respectively. Conventions for plotting follow those of Fig. 2 of the main paper.

Figure S3. Correlation between the fraction of mass collapsed ($f_{\rm coll}$,8:9) into LMACHs ($M = 10^8 - 10^9\,{\rm M}_\odot$) and the cell overdensity δ in the 114/h Mpc box, where the box is sampled by 32^3 grid-cells.
Figure S4. Correlation between the fraction of mass collapsed ($f_{\rm coll}$,9:12.5) into HMACHs ($M = 10^9 - 10^{12.5}\,{\rm M}_\odot$, where the maximum mass is roughly the mass of a cell) and the cell overdensity δ in the 114/h Mpc box, where the box is sampled by 32^3 grid-cells.
Figure S5. Same as Fig. 6(A) of the main paper, but of cells with Eulerian volume $(0.15/h\ {\rm Mpc})^3$ in 20/h Mpc box.
Figure S6. PDFs of LMACHs at given overdensity δ in cells with Eulerian volume $(3.56/h\ {\rm Mpc})^3$ in 114/h Mpc box.
Figure S7. PDFs of LMACHs at given overdensity δ in cells with Eulerian volume $(1.78/h\ {\rm Mpc})^3$ in 114/h Mpc box.
Figure S8. PDFs of HMACHs at given overdensity δ in cells with Eulerian volume $(3.56/h\ {\rm Mpc})^3$ in 114/h Mpc box.
Figure S9. PDFs of HMACHs at given overdensity δ in cells with Eulerian volume $(1.78/h\ {\rm Mpc})^3$ in 114/h Mpc box.
Figure S10. Sames as Fig. 7 of the main paper but with $V^{\rm cell} = (0.15/h)^3$ Mpc.
Figure S11. Same as Fig. 7 of the main paper but for LMACHs inside 114/h Mpc box with $V_{\rm cell} = (3.56/h)^3$ Mpc.
Figure S12. Same as Fig. 11 but with $V_{\rm cell} = (3.56/h)^3$ Mpc.
Figure S13. Same as Fig. 7 of the main paper but for HMACHs inside 114/h Mpc box with $V_{\rm cell} = (3.56/h)^3$ Mpc.
Figure S14. Same as Fig. 13 but with $V_{\rm cell} = (1.78/h)^3$ Mpc.
(http://mnras.oxfordjournals.org/lookup/suppl/doi:10.1093/mnras/stv704/-/DC1).

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the paper.

This paper has been typeset from a TEX/LATEX file prepared by the author.